

SPARSE COVARIANCE THRESHOLDING FOR HIGH-DIMENSIONAL VARIABLE SELECTION

X. Jessie Jeng and Z. John Daye

Purdue University

Abstract: In high-dimensions, many variable selection methods, such as the lasso, are often limited by excessive variability and rank deficiency of the sample covariance matrix. Covariance sparsity is a natural phenomenon in high-dimensional applications, such as microarray analysis, image processing, etc., in which a large number of predictors are independent or weakly correlated. In this paper, we propose the covariance-thresholded lasso, a new class of regression methods that can utilize covariance sparsity to improve variable selection. We establish theoretical results, under the random design setting, that relate covariance sparsity to variable selection. Real-data and simulation examples indicate that our method can be useful in improving variable selection performances.

Key words and phrases: Consistency, covariance sparsity, large p small n , random design, regression, regularization.

1. Introduction

Variable selection in high-dimensional regression is a central problem in Statistics and has stimulated much interest in the past few years. Motivation for developing effective variable selection methods in high-dimensions comes from a variety of applications, such as gene microarray analysis, image processing, etc., where it is necessary to identify a parsimonious subset of predictors to improve interpretability and prediction accuracy. In this paper, we consider the following linear model for $X = (X_1, X_2, \dots, X_p)^T$ a vector of p predictors and Y a response variable,

$$Y = X\beta^* + \epsilon, \quad (1.1)$$

where $\beta^* = (\beta_1^*, \beta_2^*, \dots, \beta_p^*)^T$ is a vector of regression coefficients and ϵ is a normal random error with mean 0 and variance σ^2 . If β_j^* is nonzero, then X_j is said to be a true variable; otherwise, it is an irrelevant variable. Further, when only a few coefficients β_j^* 's are believed to be nonzero, we refer to (1.1) as a sparse

linear model. The purpose of variable selection is to separate the true variables from the irrelevant ones based upon some observations of the model. In many applications, p can be fairly large or even larger than n . The problem of large p and small n presents a fundamental challenge for variable selection.

Recently, various methods based upon L_1 penalized least squares are proposed for variable selection. The lasso, introduced by Tibshirani (1996), is the forerunner and foundation for many of these methods. Suppose that \mathbf{y} is an $n \times 1$ vector of observed responses centered to have mean 0 and $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_p)$ is an $n \times p$ data matrix with each column \mathbf{X}_j standardized to have mean zero and variance of 1. We may reformulate the lasso as the following,

$$\hat{\beta}^{Lasso}(\lambda_n) = \arg \min_{\beta} \left\{ \beta^T \hat{\Sigma} \beta - 2\beta^T \left(\frac{1}{n} \mathbf{X}^T \mathbf{y} \right) + 2\lambda_n \|\beta\|_1 \right\}, \quad (1.2)$$

where $\hat{\Sigma} = \mathbf{X}^T \mathbf{X} / n$ is the sample covariance or correlation matrix. Consistency in variable selection for the lasso has been proved under the neighborhood stability condition in Meinshausen and Bühlmann (2006) and under the irrepresentable condition in Zhao and Yu (2006). Compared with traditional variable selection procedures, such as all subset selection, AIC, BIC, etc., the lasso has continuous solution paths and can be computed efficiently using innovative algorithms, such as the LARS in Efron, Hastie, Johnstone, and Tibshirani (2004). Since its introduction, the lasso has emerged as one of the most widely-used methods for variable selection.

In the lasso literature, data matrix \mathbf{X} is often assumed to be fixed. However, this assumption may not be realistic in high-dimensional applications, where data usually come from observational rather than experimental studies. In this paper, we assume the predictors X_1, X_2, \dots, X_p in (1.1) to be random with $E(X) = 0$ and $E(XX^T) = \Sigma = (\sigma_{ij})_{1 \leq i \leq p, 1 \leq j \leq p}$. In addition, we assume that the population covariance matrix Σ is sparse in the sense that the proportion of nonzero σ_{ij} in Σ is relatively small. Motivations for studying sparse covariance matrices come from a myriad of applications in high-dimensions, where a large number of predictors can be independent or weakly correlated with each other. For example, in gene microarray analysis, it is often reasonable to assume that genes belonging to different pathways or systems are independent or weakly correlated (Rothman, Levina, and Zhu 2009; Wagaman and Levina 2008). In these

applications, the number of nonzero covariances in Σ can be much smaller than $p(p-1)/2$, the total number of covariances.

An important component of lasso regression (1.2) is the sample covariance matrix $\hat{\Sigma}$. We note that the sample covariance matrix is rank-deficient when $p > n$. This can cause the lasso to saturate after at most n variables are selected. Moreover, the ‘large p and small n ’ scenario can cause excessive variability of sample covariances between the true and irrelevant variables. This deteriorates the ability of the lasso to separate true variables from irrelevant ones. More specifically, a sufficient and almost necessary condition for the lasso to be variable selection consistent is derived in Zhao and Yu (2006), which they call the irrepresentable condition. It poses constraint on the inter-connectivity between the true and irrelevant variables in the following way. Let $S = \{j \in \{1, \dots, p\} \mid \beta_j^* \neq 0\}$ and $C = \{1, 2, \dots, p\} - S$, such that S is the collection of true variables and C is the complement of S that is composed of the irrelevant variables. Assume that the cardinality of S is s ; in other words, there are s true variables and $p - s$ irrelevant ones. Further, let \mathbf{X}_S and \mathbf{X}_C be sub-data matrices of \mathbf{X} that contain the observations of the true and irrelevant variables, respectively. Define $\hat{\mathcal{I}} = |\hat{\Sigma}_{CS}(\hat{\Sigma}_{SS})^{-1} \text{sgn}(\beta_S^*)|$, where $\hat{\Sigma}_{CS} = \mathbf{X}_C^T \mathbf{X}_S / n$ and $\hat{\Sigma}_{SS} = \mathbf{X}_S^T \mathbf{X}_S / n$. We refer to $\hat{\mathcal{I}}$ as the *sample irrepresentable index*. It can be interpreted as representing the amount of inter-connectivity between the true and irrelevant variables. In order for lasso to select the true variables consistently, irrepresentable condition requires $\hat{\mathcal{I}}$ to be bounded from above, that is $\hat{\mathcal{I}} < 1 - \epsilon$ for some $\epsilon \in (0, 1)$, entry-wise. Clearly, excessive variability of the sample covariance matrix induced by large p and small n can cause $\hat{\mathcal{I}}$ to exhibit large variation that makes the irrepresentable condition less likely to hold. These inadequacies motivate us to consider alternatives to the sample covariance matrix to improve variable selection for the lasso in high-dimensions.

Next, we provide some insight on how the sparsity of the population covariance matrix can influence variable selection for the lasso. Under random design assumption on X , the inter-connectivity between the true and irrelevant variables can be stated in terms of their population variances and covariances. Let Σ_{CS} be the covariance matrix between the irrelevant variables and true variables and Σ_{SS} the variance-covariance matrix of the true variables. We define the *popula-*

tion irrerepresentable index as $\mathcal{I} = |\Sigma_{CS}\Sigma_{SS}^{-1}\text{sgn}(\beta_S^*)|$. Intuitively, the sparser the population covariances Σ_{CS} and Σ_{SS} are, or the sparser Σ is, the more likely that $\mathcal{I} < 1 - \epsilon$, entry-wise. This property, however, does not automatically trickle down to the sample irrerepresentable index $\hat{\mathcal{I}}$, due to its excessive variability. When Σ_{CS} and Σ_{SS} are known a priori to be sparse and $\mathcal{I} < 1 - \epsilon$, entry-wise, some regularization on the covariance can be used to reduce the variabilities of $\hat{\Sigma}$ and $\hat{\mathcal{I}}$ and allow the irrerepresentable condition to hold more easily for $\hat{\mathcal{I}}$. Furthermore, the sample covariance matrix $\hat{\Sigma} = \mathbf{X}^T\mathbf{X}/n$ is obviously non-sparse; and imposing sparsity on $\hat{\Sigma}$ has the benefit of sometimes increasing the rank of the sample covariance matrix.

We use an example to demonstrate how rank deficiency and excessive variability of the sample covariance matrix $\hat{\Sigma}$ can compromise the performance of the lasso for large p and small n . Suppose there are 40 variables ($p = 40$) and $\Sigma = I_p$ (I_p is the $p \times p$ identity matrix). Since all variables are independent of each other, the population irrerepresentable index clearly satisfies $\mathcal{I} < 1 - \epsilon$, entry-wise. Further, we let $\beta_j^* = 2$, for $1 \leq j \leq 10$, and $\beta_j^* = 0$, for $11 \leq j \leq 40$. The error standard deviation σ is set to be about 6.3 to have a signal-to-noise ratio of approximately 1. The lasso, in general, does not take into consideration the structural properties of the model, such as the sparsity or the orthogonality of Σ in this example. One way to take advantage of the orthogonality of Σ is to replace $\hat{\Sigma}$ in (1.2) by I_p , which leads to the univariate soft thresholding (UST) estimates $\hat{\beta}_j^{UST} = \text{sgn}(r_j)(|r_j| - \lambda)^+$, where $r_j = \mathbf{X}_j^T\mathbf{Y}/n$ for $1 \leq j \leq p$. We compare the performances of the lasso and UST over various sample sizes ($5 \leq n \leq 250$) using the variable selection measure G . G is defined as the geometric mean between sensitivity, (no. of true positives)/ s , and specificity, $1 - (\text{no. of false positives})/(p - s)$ (Tibshirani, Saunders, Rosset, Zhu, and Knight 2005; Chong and Jun 2005; Kubat, Holte, and Matwin 1998). G varies between 0 and 1. Larger G indicates better selection with a larger proportion of variables classified correctly.

Figure 1 plots the median G based on 200 replications for the lasso and UST against sample sizes. For each replication, λ is determined ex post facto by the optimal G in order to avoid stochastic errors from tuning parameter estimation, such as by using cross-validation. It is clear from Figure 1 that, when $n > 20$,

lasso slightly outperforms UST; when $n < 20$, the performance of lasso starts to deteriorate precipitously, whereas the performance of UST declines at a much slower pace and starts to outperform lasso. This example suggests that when p is large and n is relatively small, sparsity of Σ can be used to enhance variable selection.

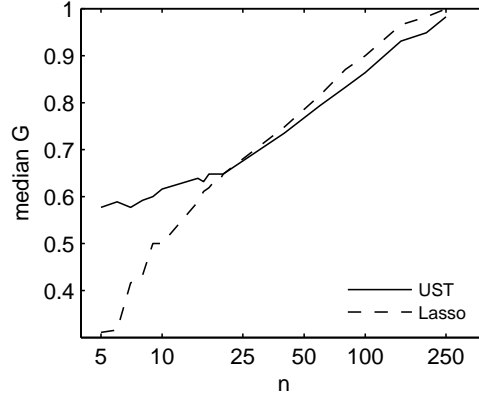


Figure 1: Median G (from $n=5$ to $n=250$) for illustrating example based upon 200 replications.

The discussions above motivate us to consider improving the performance of the lasso by applying regularization to the sample covariance matrix $\hat{\Sigma}$. A good sparse covariance-regularizing operator on $\hat{\Sigma}$ should satisfy the following properties:

1. The operator *stabilizes* $\hat{\Sigma}$.
2. The operator can *increase the rank* of $\hat{\Sigma}$.
3. The operator *utilizes the underlying sparsity* of the covariance matrix.

The first and second properties are obviously useful and have been explored in the literature. For example, the elastic net, introduced in Zou and Hastie (2005), replaces $\hat{\Sigma}$ by $\hat{\Sigma}_{EN} = (\hat{\Sigma} + \lambda_2 I)/(1 + \lambda_2)$ in (1.2), where $\lambda_2 > 0$ is a tuning parameter. $\hat{\Sigma}_{EN}$ can be more stable and have higher rank than $\hat{\Sigma}$ but is non-sparse. Nonetheless, in many applications, utilizing the underlying sparsity may be more crucial in improving the lasso when data is scarce, such as under the large p and small n scenario.

Recently, various regularization methods have been proposed in the literature for estimating high-dimensional variance-covariance matrices. Some examples include tapering proposed by Furrer and Bengtsson (2007), banding by Bickel and Levina (2008b), thresholding by Bickel and Levina (2008a) and El Karoui (2008), and generalized thresholding by Rothman, Levina, and Zhu (2009). We note that covariance thresholding operators can satisfy all three properties outlined in the previous paragraph; in particular, they can generate sparse covariance estimates to accommodate for the covariance sparsity assumption. In this paper, we propose to apply covariance-thresholding on the sample covariance matrix $\hat{\Sigma}$ in (1.2) to stabilize and improve the performances of the lasso. We call this procedure the *covariance-thresholded lasso*. We establish theoretical results that relate the sparsity of the covariance matrix with variable selection and compare them to those of the lasso. Simulation and real-data examples are reported. Our results suggest that covariance-thresholded lasso can improve upon the lasso, adaptive lasso, and elastic net, especially when Σ is sparse, n is small, and p is large. Even when the underlying covariance is non-sparse, covariance-thresholded lasso is still useful in providing robust variable selection in high-dimensions.

Witten and Tibshirani (2009) has recently proposed the scout procedure, that applies regularization to the inverse covariance or precision matrix. We note that this is quite different from the covariance-thresholded lasso that regularizes the sample covariance matrix $\hat{\Sigma}$ directly. Furthermore, the scout penalizes using the matrix norm $\|\Theta_{\mathbf{X}\mathbf{X}}\|_p^p = \sum_{ij} |\theta_{ij}|^p$, where Θ is an estimate of Σ^{-1} , whereas the covariance-thresholded lasso regularizes individual covariances $\hat{\sigma}_{ij}$ directly. In our results, we will show that the scout is potentially very similar to the elastic net and that the covariance-thresholded lasso can often outperform the scout in terms of variable selection for $p > n$.

The rest of the paper is organized as follows. In Section 2, we present covariance-thresholded lasso in detail and a modified LARS algorithm for our method. We discuss a generalized class of covariance-thresholding operators and explain how covariance-thresholding can stabilize the LARS algorithm for the lasso. In Section 3, we establish theoretical results on variable selection for the covariance-thresholded lasso. The effect of covariance sparsity on variable selection is especially highlighted. In Section 4, we provide simulation results

of covariance-thresholded lasso at $p > n$, and, in Section 5, we compare the performances of covariance-thresholded lasso with those of the lasso, adaptive lasso, and elastic net using 3 real-data sets. Section 6 concludes with further discussions and implications.

2. The Covariance-Thresholded Lasso

Suppose that the response \mathbf{y} is centered and each column of the data matrix \mathbf{X} is standardized, as in the lasso (1.2). We define the covariance-thresholded lasso estimate as

$$\hat{\beta}^{CT-Lasso}(\nu, \lambda_n) = \arg \min_{\beta} \left\{ \beta^T \hat{\Sigma}_{\nu} \beta - 2\beta^T \left(\frac{1}{n} \mathbf{X}^T \mathbf{y} \right) + 2\lambda_n \|\beta\|_1 \right\}, \quad (2.3)$$

where $\hat{\Sigma}_{\nu} \equiv [\hat{\sigma}_{ij}^{\nu}]$, $\hat{\sigma}_{ij}^{\nu} = s_{\nu}(\hat{\sigma}_{ij})$, $\hat{\sigma}_{ij} = \sum_{k=1}^n X_{ki} X_{kj} / n$, and $s_{\nu}(\cdot)$ is a pre-defined covariance-thresholding operator with $0 \leq \nu < 1$. If the identity function is used as the covariance-thresholding operator, that is $s_{\nu}(x) = x$ for any x , then $\hat{\beta}^{CT-Lasso}(\nu, \lambda_n) = \hat{\beta}^{Lasso}$.

2.1. Sparse Covariance-thresholding Operators

We consider a generalized class of covariance-thresholding operators $s_{\nu}(\cdot)$ introduced in Rothman, Levina, and Zhu (2009). These operators satisfy the following properties,

$$s_{\nu}(\hat{\sigma}_{ij}) = 0 \text{ for } |\hat{\sigma}_{ij}| \leq \nu, \quad |s_{\nu}(\hat{\sigma}_{ij})| \leq |\hat{\sigma}_{ij}|, \quad |s_{\nu}(\hat{\sigma}_{ij}) - \hat{\sigma}_{ij}| \leq \nu. \quad (2.4)$$

The first property enforces sparsity for covariance estimation; the second allows shrinkage of covariances; and the third limits the amount of shrinkage. These operators satisfy the desired properties outlined in the Introduction for sparse covariance-regularizing operators and represent a wide spectrum of thresholding procedures that can induce sparsity and stabilize the sample covariance matrix. In this paper, we will consider the following covariance-thresholding operators for $\hat{\sigma}_{ij}$ when $i \neq j$.

$$1. \text{ Hard thresholding: } s_{\nu}^{\text{Hard}}(\hat{\sigma}_{ij}) = \hat{\sigma}_{ij} 1(|\hat{\sigma}_{ij}| > \nu). \quad (2.5)$$

$$2. \text{ Soft thresholding: } s_{\nu}^{\text{Soft}}(\hat{\sigma}_{ij}) = \text{sgn}(\hat{\sigma}_{ij})(|\hat{\sigma}_{ij}| - \nu)^+. \quad (2.6)$$

$$3. \text{ Adaptive thresholding: For } \gamma \geq 0,$$

$$s_{\nu}^{\text{Adapt}}(\hat{\sigma}_{ij}) = \text{sgn}(\hat{\sigma}_{ij})(|\hat{\sigma}_{ij}| - \nu^{\gamma+1} |\hat{\sigma}_{ij}|^{-\gamma})^+. \quad (2.7)$$

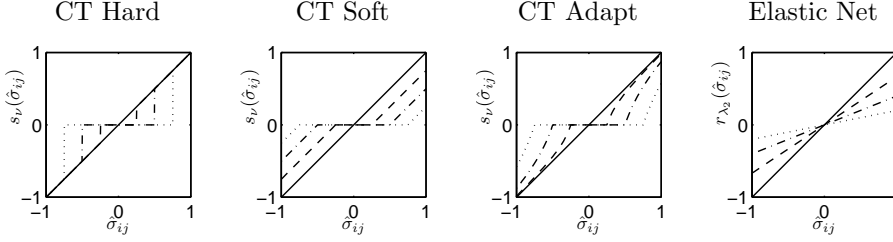


Figure 2: Hard, soft, adaptive ($\gamma=2$) sparse covariance-thresholding operators with ν varying over 0 (solid), 0.25 (dashed), 0.5 (dot-dashed), and 0.75 (dotted); and elastic net covariance-regularizing operator with λ_2 varying over 0 (solid), 0.5 (dashed), 1.5 (dot-dashed), and 4 (dotted).

The above operators are used in Rothman, Levina, and Zhu (2009) for estimating variance-covariance matrices, and it is easy to check that they satisfy the properties in (2.4).

In Figure 2, we depict the sparse covariance-thresholding operators (2.5-2.7) for varying ν . Hard thresholding presents a discontinuous thresholding of covariances, whereas soft thresholding offers continuous shrinkage. Adaptive thresholding presents less regularization on covariances with large magnitudes than soft thresholding.

Figure 2 further includes the elastic net covariance-regularizing operator, $r_{\lambda_2}(\hat{\sigma}_{ij}) = (\hat{\sigma}_{ij} + \lambda_2)/(1 + \lambda_2)$ for $i \neq j$. Apparently, this operator is non-sparse and does not satisfy the first property in (2.4). In particular, we see that the elastic net penalizes covariances with large magnitudes more severely than those with small magnitudes. In some situations, this has the benefit of alleviating multicollinearity as it shrinks covariances of highly correlated variables. However, under high-dimensionality and when much of the random perturbation of the covariance matrix arises from small but numerous covariances, the elastic net in attempting to control these variabilities may inadvertently penalize covariances with large magnitudes severely, which may introduce large bias in estimation and compromise the performance of the elastic net under some scenarios.

2.2. Computations

The lasso solution paths are shown to be piecewise linear in Efron, Hastie, Johnstone, and Tibshirani (2004) and Rosset and Zhu (2007). This property allows Efron, Hastie, Johnstone, and Tibshirani (2004)

to propose the efficient LARS algorithm for the lasso. Likewise, in this section, we propose a piecewise-linear algorithm for the covariance-thresholded lasso.

We note that the loss function $\beta^T \hat{\Sigma}_\nu \beta - 2\beta^T \mathbf{X}^T \mathbf{y}/n$ in (2.3) can sometimes be non-convex since $\hat{\Sigma}_\nu$ may possess negative eigenvalues for some ν . This usually may occur for intermediary values of ν , as $\hat{\Sigma}_\nu$ is at least semi-positive definite for ν close to 0 or 1. Furthermore, we note that the penalty $2\lambda_n \|\beta\|_1$ is a convex function and dominates in (2.3) for λ_n large. Intuitively, this means that the optimization problem for covariance-thresholded lasso is almost convex for β sparse. This is stated conservatively in the following theorem by using second-order condition from nonlinear programming (McCormick 1976).

Theorem 2.1 *Let ν be fixed. If $\hat{\Sigma}_\nu$ is semi-positive definite, the covariance-thresholded lasso solutions $\hat{\beta}^{CT-Lasso}(\nu, \lambda_n)$ for (2.3) are piecewise linear with respect to λ_n . If $\hat{\Sigma}_\nu$ possesses negative eigenvalues, a set of covariance-thresholded lasso solutions, which may be local minima for (2.3) under strict complementarity, is piecewise linear with respect to λ_n for $\lambda_n \geq \lambda^*$, where $\lambda^* = \min\{\lambda > 0 : \text{sub-matrix } (\hat{\Sigma}_\nu)_\mathcal{A} \text{ remains positive definite for } \mathcal{A} = \{j : \hat{\beta}_j^{CT-Lasso}(\nu, \lambda) \neq 0\}\}$*

The proof for Theorem 2.1 is outlined in Appendix 7.6. Strict complementarity, described in Appendix 7.6, is a technical condition that allows the second-order condition to be more easily interpreted and usually holds with high probability. We note that, when $\hat{\Sigma}_\nu$ has negative eigenvalues, the solution $\hat{\beta}^{CT-Lasso}(\nu, \lambda_n)$ is global if $|\mathbf{x}_j^T \mathbf{y}/n| < \lambda_n$ for all $j \notin \mathcal{A}_n = \{j : \hat{\beta}_j^{CT-Lasso}(\nu, \lambda_n) \neq 0\}$ and $(\hat{\Sigma}_\nu)_{\mathcal{A}_n}$ is positive definite. Theorem 2.1 suggests that piecewise linearity of the covariance-thresholded lasso solution path sometimes may not hold for some ν when λ_n is small, even if a solution may well exist. This restricts the sets of tuning parameters (ν, λ_n) for which we can compute the solutions of covariance-thresholded lasso efficiently using a LARS-type algorithm. We note that the elastic net does not suffer from a potentially non-convex optimization. However, as we will demonstrate in Figure 3 of Section 4, covariance-thresholded lasso with restricted sets of (ν, λ_n) is, nevertheless, rich enough to dominate the elastic net in many situations.

Theorem 2.1 establishes that a set of covariance-thresholded lasso solutions are piecewise linear. This further provides us with an efficient modified LARS

algorithm for computing the covariance-thresholded lasso. Let

$$(\hat{c}_\nu)_j = \frac{1}{n} \mathbf{X}_j^T \mathbf{y} - (\hat{\Sigma}_\nu)_j^T \beta \quad (2.8)$$

be estimates for the covariate-residual correlations c_j . Further, we denote the minimum eigenvalue of A as $\Lambda_{\min}(A)$. The covariance-thresholded lasso can be computed with the following algorithm.

ALGORITHM: Covariance-thresholded LARS

1. Initialize $\hat{\Sigma}_\nu$ such that $\hat{\sigma}_{ij}^\nu = s_\nu(\hat{\sigma}_{ij})$, $\beta = 0$, and $\hat{\mathbf{c}}_\nu = \frac{1}{n} \mathbf{X}^T \mathbf{y}$. Let $\mathcal{A} = \arg \max_j |(\hat{c}_\nu)_j|$, $\hat{C} = \max |(\hat{\mathbf{c}}_\nu)_\mathcal{A}|$, $\gamma_\mathcal{A} = \text{sgn}((\hat{\mathbf{c}}_\nu)_\mathcal{A})$, $\gamma_{\mathcal{A}^c} = 0$, and $\mathbf{a} = (\hat{\Sigma}_\nu)^T \gamma$.
2. Let $\delta_1 = \min_{j \in \mathcal{A}}^+ \{-\frac{\beta_j}{\gamma_j}\}$ and $\delta_2 = \min_{j \in \mathcal{A}^c}^+ \{\frac{\hat{C} - (\hat{c}_\nu)_j}{a_i - a_j}, \frac{\hat{C} + (\hat{c}_\nu)_j}{a_i + a_j}\}$ for any $i \in \mathcal{A}$, where \min^+ is taken only over positive elements.
3. Let $\delta = \min(\delta_1, \delta_2)$, $\beta \leftarrow \beta + \delta \gamma$, $\hat{\mathbf{c}}_\nu \leftarrow \hat{\mathbf{c}}_\nu - \delta \mathbf{a}$, and $\hat{C} = \max_{j \in \mathcal{A}} |(\hat{c}_\nu)_j|$.
4. If $\delta = \delta_1$, remove the variable hitting 0 at δ from \mathcal{A} . If $\delta = \delta_2$, add the variable first attaining equality at δ to \mathcal{A} .
5. Compute the new direction, $\gamma_\mathcal{A} = (\hat{\Sigma}_\nu)_\mathcal{A}^{-1} \text{sgn}(\beta_\mathcal{A})$ and $\gamma_{\mathcal{A}^c} = 0$, and let $\mathbf{a} = (\hat{\Sigma}_\nu)^T \gamma$.
6. Repeat steps 2-5 until $\min_{j \in \mathcal{A}} |(\hat{c}_\nu)_j| < 0$ or $\Lambda_{\min}((\hat{\Sigma}_\nu)_\mathcal{A}) \leq 0$.

The covariate-residual correlations c_j are the most crucial for computing the solution paths. It determines the variable to be included at each step and relates directly to the tuning parameter λ_n . In the original LARS for the lasso, c_j is estimated as $\mathbf{X}_j^T \mathbf{y} / n - \hat{\Sigma}_j^T \beta$, which uses the sample covariance matrix $\hat{\Sigma}$ without thresholding. In covariance-thresholded LARS, $(\hat{c}_\nu)_j$ is defined using the covariance-thresholded estimate $(\hat{\Sigma}_\nu)_j^T = (\hat{\sigma}_{1j}^\nu, \hat{\sigma}_{2j}^\nu, \dots, \hat{\sigma}_{pj}^\nu)$, which may contain many zeros. We note that, in (2.8), zero-valued covariances $\hat{\sigma}_{ij}^\nu$ have the effect of essentially removing the associated coefficients from β , providing parsimonious estimates for c_j . This allows covariance-thresholded LARS to estimate c_j in a more stable way than the LARS. It is clear that covariance-thresholded LARS presents an advantage if population covariance is sparse. On the other hand, if the covariance is non-sparse, covariance-thresholded LARS can still outperform the LARS when the sample size is small or the data are noisy. This is because

parsimonious estimates $(\hat{c}_\nu)_j$ of c_j can be more robust against random variability of the data.

Moreover, consider computing the direction of the solution paths in Step 5, which is used for updating $(\hat{c}_\nu)_j$. LARS for the lasso updates new directions with $(\hat{\Sigma})_{\mathcal{A}}^{-1} \text{sgn}(\beta_{\mathcal{A}})$, whereas covariance-thresholded LARS uses $\gamma_{\mathcal{A}} = (\hat{\Sigma}_\nu)_{\mathcal{A}}^{-1} \text{sgn}(\beta_{\mathcal{A}})$. Apparently, covariance-thresholded LARS can exploit potential covariance sparsity to improve and stabilize estimates of the directions of the solution paths. In addition, the LARS for the lasso can stop early before all true variables S can be considered if $\hat{\Sigma}_{\mathcal{A}}$ is rank deficient at an early stage when sample size is limited. Covariance-thresholding can mitigate this problem by proceeding further with properly chosen values of ν . For example, when $\nu \rightarrow 1$, $\hat{\Sigma}_\nu$ converges towards the identity matrix \mathbf{I} , which is full-ranked.

3. Theoretical Results on Variable Selection

In this section, we derive sufficient conditions for covariance-thresholded lasso to be consistent in selecting the true variables. We relate covariance sparsity with variable selection and demonstrate the pivotal role that covariance sparsity plays in improving variable selection under high-dimensionality. Furthermore, variable selection results for the lasso under the random design are derived and compared with those of the covariance-thresholded lasso. We show that the covariance-thresholded lasso, by utilizing covariance sparsity through a properly chosen thresholding level ν , can improve upon the lasso in terms of variable selection.

For simplicity, we assume that a solution for (2.3) exists and denote the covariance-thresholded lasso estimate $\hat{\beta}^{CT-Lasso}(\nu, \lambda_n)$ by $\hat{\beta}^\nu$ in this section. Further, we let $\text{supp}(\beta) = \{j : \beta_j \neq 0\}$ represent the collection of indices of nonzero coefficients. We say that the covariance-thresholded lasso estimate $\hat{\beta}^\nu$ is variable selection consistent if $P(\text{supp}(\hat{\beta}^\nu) = \text{supp}(\beta^*)) \rightarrow 1$, as $n \rightarrow \infty$. In addition, we say that $\hat{\beta}^\nu$ is sign consistent if $P(\text{sgn}(\hat{\beta}^\nu) = \text{sgn}(\beta^*)) \rightarrow 1$, as $n \rightarrow \infty$, where $\text{sgn}(t) = -1, 0, 1$ when $t < 0$, $t = 0$ and $t > 0$, respectively (Zhao and Yu 2006). Obviously, sign consistency is a stronger property and implies variable selection consistency.

We introduce two quantities to characterize the sparsity of Σ that plays a pivotal role in the performance of covariance-thresholded lasso. Recall that S

and C are collections of the true and irrelevant variables, respectively. Define

$$d_{SS}^* = \max_{i \in S} \sum_{j \in S} 1(\sigma_{ij} \neq 0) \quad \text{and} \quad d_{CS}^* = \max_{i \in C} \sum_{j \in S} 1(\sigma_{ij} \neq 0). \quad (3.9)$$

d_{SS}^* ranges between 1 and s . When $d_{SS}^* = 1$, all pairs of the true variables are orthogonal. When $d_{SS}^* = s$, there are at least one variable correlated with all other variables. Similarly, d_{CS}^* is between 0 and s . When $d_{CS}^* = 0$, the true and irrelevant variables are orthogonal to each other, and, when $d_{CS}^* = s$, some irrelevant variables are correlated with all the true variables. The values of d_{SS}^* and d_{CS}^* represent the sparsity of covariance sub-matrices for the true variables and between the irrelevant and true variables, respectively. We have not specified the sparsity of the sub-matrix for the irrelevant variables themselves. It will be clear later that it is the structure of Σ_{SS} and Σ_{CS} instead of Σ_{CC} that plays the pivotal role in variable selection. We note that d_{SS}^* and d_{CS}^* are related to another notion of sparsity used in Bickel and Levina (2008a) to define the class of matrices $\{\Sigma : \sigma_{ii} \leq M, \sum_{j=1}^p 1(\sigma_{ij} \neq 0) \leq c_0(p) \text{ for } 1 \leq i \leq p\}$, for M given and $c_0(p)$ a constant depending on p . We use the specific quantities d_{SS}^* and d_{CS}^* in (3.9) in order to provide easier presentation of our results for variable selection. Our results in this section can be applied to more general characterizations of sparsity, such as in Bickel and Levina (2008a).

In this paper, we employ two different types of matrix norms. For an arbitrary matrix $A = [A_{ij}]$, the infinity norm is defined as $\|A\|_\infty = \max_i \sum_j |A_{ij}|$, and the spectral norm is defined as $\|A\| = \max_{x: \|x\| \leq 1} \|Ax\| = \Lambda_{\max}(A)$. We use $\Lambda_{\max}(A)$ and $\Lambda_{\min}(A)$ to represent, respectively, the largest and smallest eigenvalues of A .

3.1. Sign Consistency of Covariance-thresholded Lasso

We develop sign consistency results for covariance-thresholded lasso. Proofs for the results are presented in the Appendix.

We first provide conditions for the covariance-thresholded lasso estimate $\hat{\beta}^\nu$ to have the same signs as the true coefficients β^* under the fixed design assumption. Let $\bar{\rho} = \max_{j \in S} |\beta_j^*|$ and $\underline{\rho} = \min_{j \in S} |\beta_j^*|$.

Lemma 3.1 *Suppose that the data matrix \mathbf{X} is fixed and ν is given. Then,*

$\text{sgn}(\hat{\beta}^\nu) = \text{sgn}(\beta^*)$ if

$$\Lambda_{\min}(\hat{\Sigma}_{SS}^\nu) > 0, \quad (3.10)$$

$$\left\| \hat{\Sigma}_{CS}^\nu (\hat{\Sigma}_{SS}^\nu)^{-1} \right\|_\infty \left(\left\| \frac{1}{n} \mathbf{X}_S^T \epsilon \right\|_\infty + s\nu\bar{\rho} + \lambda_n \right) + s\nu\bar{\rho} + \left\| \frac{1}{n} \mathbf{X}_C^T \epsilon \right\|_\infty \leq \lambda_n, \quad (3.11)$$

and

$$\left\| (\hat{\Sigma}_{SS}^\nu)^{-1} \right\|_\infty \left(\left\| \frac{1}{n} \mathbf{X}_S^T \epsilon \right\|_\infty + s\nu\bar{\rho} + \lambda_n \right) < \underline{\rho}. \quad (3.12)$$

The above (3.10), (3.11), and (3.12) are derived from the Karush-Kuhn-Tucker (KKT) conditions for the optimization problem presented in (2.3) when the solution, which may be a local minimum, exists. Following the arguments in Zhao and Yu (2006) and Wainwright (2006), these conditions are almost necessary for $\hat{\beta}^\nu$ to have the correct signs. The condition (3.10) is needed for (3.11) and (3.12) to be valid. That is, the conditions (3.11) and (3.12) are ill-defined if $\hat{\Sigma}_{SS}^\nu$ is singular.

Assume the random design setting so that \mathbf{X} is drawn from some distribution with population covariance Σ . We demonstrate how the sparsity of Σ_{SS} and the procedure of covariance-thresholding work together to ensure that the condition (3.10) is satisfied. We impose the following moment conditions on the random predictors X_1, \dots, X_p :

$$EX_j = 0, \quad EX_j^{2d} \leq d!M^d, \quad 1 \leq j \leq p, \quad (3.13)$$

for some constant $M > 0$ and $d \in \mathbb{N}$. Assume that

$$\Lambda_{\min}(\Sigma_{SS}) > 0 \quad (3.14)$$

and d_{SS}^* , s , and n satisfy

$$d_{SS}^* \sqrt{\log s} / \sqrt{n} \rightarrow 0. \quad (3.15)$$

We have the following lemma.

Lemma 3.2 *Let $\nu = C\sqrt{\log s}/\sqrt{n}$ for some constant $C > 0$. Under the conditions (3.13), (3.14), and (3.15),*

$$P\left(\Lambda_{\min}(\hat{\Sigma}_{SS}^\nu) > 0\right) \rightarrow 1. \quad (3.16)$$

The rate of convergence for (3.16) depends on the rate of convergence for (3.15). It is clear that the smaller d_{SS}^* (or the sparser Σ_{SS}) is, the faster (3.15), as well as (3.16), converges. Equivalently, for sample size n fixed, the smaller d_{SS}^* is, the larger the probability that $\Lambda_{\min}(\hat{\Sigma}_{SS}^\nu) > 0$. In other words, covariance-thresholding can help to fix potential rank deficiency of $\hat{\Sigma}_{SS}$ when Σ_{SS} is sparse. In the special case when $\Sigma_{SS} = I_p$ and $d_{SS}^* = 1$, it can be shown that $\hat{\Sigma}_{SS}^\nu$ is asymptotically positive definite provided that $s = o(\exp(n))$.

Next, we investigate the remaining two conditions (3.11) and (3.12) in Lemma 3.1. For (3.11) and (3.12) to hold with probability going to 1, additional assumptions including the irrepresentable condition need to be imposed. Since the data matrix \mathbf{X} is assumed to be random, the original irrepresentable condition needs to be stated in terms of the population covariance matrix Σ as follows,

$$\|\Sigma_{CS}(\Sigma_{SS})^{-1}\|_\infty \leq 1 - \epsilon, \quad (3.17)$$

for some $0 < \epsilon < 1$. We note that the original irrepresentable condition in Zhao and Yu (2006) also involves the signs of β_S^* . To simplify presentation, we use the stronger condition (3.17) instead. Obviously, (3.17) does not directly imply that $\|\hat{\Sigma}_{CS}^\nu(\hat{\Sigma}_{SS}^\nu)^{-1}\|_\infty \leq 1 - \epsilon$. The next lemma establishes the asymptotic behaviors of $\|(\hat{\Sigma}_{SS}^\nu)^{-1}\|_\infty$ and $\|\hat{\Sigma}_{CS}^\nu(\hat{\Sigma}_{SS}^\nu)^{-1}\|_\infty$. Let $\bar{D} = \|(\Sigma_{SS})^{-1}\|_\infty$. Assume

$$\bar{D}d_{SS}^*\sqrt{\log(p-s)}/\sqrt{n} \rightarrow 0, \quad (3.18)$$

$$\bar{D}^2d_{CS}^*d_{SS}^*\sqrt{\log(p-s)}/\sqrt{n} \rightarrow 0. \quad (3.19)$$

Lemma 3.3 *Suppose that $p - s > s$ and $\nu = C\sqrt{\log(s(p-s))}/\sqrt{n}$ for some constant $C > 0$. Under conditions (3.13), (3.14), (3.17), (3.18), and (3.19),*

$$P\left(\left\|\left(\hat{\Sigma}_{SS}^\nu\right)^{-1}\right\|_\infty \leq \bar{D}\right) \rightarrow 1, \quad (3.20)$$

$$P\left(\left\|\hat{\Sigma}_{CS}^\nu(\hat{\Sigma}_{SS}^\nu)^{-1}\right\|_\infty \leq 1 - \frac{\epsilon}{2}\right) \rightarrow 1. \quad (3.21)$$

The above lemma indicates that with a properly chosen thresholding parameter ν and sample size depending on covariance-sparsity quantities d_{SS}^* and d_{CS}^* , both $\|(\hat{\Sigma}_{SS}^\nu)^{-1}\|_\infty$ and $\|\hat{\Sigma}_{CS}^\nu(\hat{\Sigma}_{SS}^\nu)^{-1}\|_\infty$ behave as their population counterparts $\|(\Sigma_{SS})^{-1}\|_\infty$ and $\|\Sigma_{CS}\Sigma_{SS}^{-1}\|_\infty$, asymptotically. Again, the influence of the sparsity of Σ on $\|(\hat{\Sigma}_{SS}^\nu)^{-1}\|_\infty$ and $\|\hat{\Sigma}_{CS}^\nu(\hat{\Sigma}_{SS}^\nu)^{-1}\|_\infty$ is shown through d_{CS}^* and d_{SS}^* .

Asymptotically, the smaller d_{CS}^* and d_{SS}^* are, the faster (3.20) and (3.21) converge. Or equivalently, for sample size n fixed, the smaller d_{CS}^* and d_{SS}^* are, the larger the probabilities in (3.20) and (3.21) are. In the special case when $d_{CS}^* = 0$ or Σ_{CS} is a zero matrix, condition (3.19) is always satisfied.

Finally, we are ready to state the sign consistency result for $\hat{\beta}^\nu$. With the help of Lemmas 1–3 stated above, the only issue left is to show the existence of a proper λ_n such that (3.11) and (3.12) hold with probability going to 1. One more condition is needed. We assume that

$$\bar{D}\bar{\rho}s\sqrt{\log(p-s)}/(\underline{\rho}\sqrt{n}) \rightarrow 0. \quad (3.22)$$

Theorem 3.2 *Suppose that $p - s > s$, $\nu = C\sqrt{\log(s(p-s))}/\sqrt{n}$ for some constant $C > 0$, and λ_n is chosen such that $\lambda_n \rightarrow 0$,*

$$\sqrt{n}\lambda_n/(s\bar{\rho}\sqrt{\log(p-s)}) \rightarrow \infty, \quad \text{and} \quad \bar{D}\lambda_n/\underline{\rho} \rightarrow 0. \quad (3.23)$$

Then, under conditions (3.13), (3.14), (3.17), (3.19), and (3.22),

$$P\left(\text{sgn}(\hat{\beta}^\nu) = \text{sgn}(\beta^*)\right) \rightarrow 1. \quad (3.24)$$

We note that the assumption $p - s > s$ is natural for high-dimensional sparse models, which usually have a large number of irrelevant variables. This assumption effects the conditions (3.19) and (3.22) as well as choices of ν and λ_n . When $p - s < s$, that is a non-sparse linear model is assumed, the conditions for $\hat{\beta}^\nu$ to be sign consistent need to be modified by choosing ν as $\nu = C\sqrt{\log s}/\sqrt{n}$ and replacing $\sqrt{\log(p-s)}$ by $\sqrt{\log s}$ in conditions (3.19), (3.22), and (3.23).

It is possible to establish the convergence rate for the probability in (3.24) more explicitly. For simplicity of presentation, we provide such a result under a special case in the following theorem.

Theorem 3.3 *Suppose that conditions (3.13), (3.14), and (3.17) hold and \bar{D} , $\underline{\rho}$, and $\bar{\rho}$ are constants. Let $\lambda_n = n^{-c}$, $\nu = n^{-c_1}$, $d_S^* = \max\{d_{SS}^*, d_{CS}^*\} = n^{c_2}$, $s = n^{c_3}$, and $\log p = o(n^{1-2c} + n(n^{-c_2} - n^{-c_1})^2)$, where c, c_1, c_2 , and c_3 are positive constants such that $c < 1/2$, $c_1 < 1/2$, $c_2 < 1/4$, $c_2 < c_3$, and $c_3 + c < c_1$. Then,*

$$P\left(\text{sgn}(\hat{\beta}^\nu) = \text{sgn}(\beta^*)\right) \geq 1 - O\left(\exp(-\alpha_1 n^{1-2c})\right) - O\left(\exp(-\alpha_2 n(n^{-2c_2} - n^{-c_1})^2)\right) \rightarrow 1, \quad (3.25)$$

where α_1 and α_2 are some positive constants depending on ϵ , \bar{D} , M and $\underline{\rho}$.

The proof of Theorem 3.3, which we omit, is similar to that of Theorem 3.2. We note that the conditions on dimension parameters in Theorem 3.2 are now expressed in the convergence rate of (3.25). It is clear that the smaller d_S^* is, the larger the probability is in (3.25).

3.2. Comparison with the Lasso

We compare sign consistency results of covariance-thresholded lasso with those of the lasso. By choosing $\nu = 0$, the covariance-thresholded lasso estimate $\hat{\beta}^\nu$ can be reduced to the lasso estimate $\hat{\beta}^0$. Results on sign consistency of the lasso have been established in the literature (Zhao and Yu (2006), Meinshausen and Bühlmann (2006), Wainwright (2006)). To facilitate comparison, we restate sign consistency results for $\hat{\beta}^0$ in the same way that we presented results for $\hat{\beta}^\nu$ in Section 3.1. The proofs, which we omit, for sign consistency of $\hat{\beta}^0$ is similar to those for $\hat{\beta}^\nu$.

First, assuming fixed design, we have the sufficient and almost necessary conditions for $\text{sgn}(\hat{\beta}^0) = \text{sgn}(\beta^*)$ as in (3.10)-(3.12) with $\nu = 0$.

Next, we assume the random design. Analogous to Lemma 3.2, the sufficient conditions for $P(\Lambda_{\min}(\hat{\Sigma}_{SS}) > 0) \rightarrow 1$ are (3.13), (3.14), and

$$s\sqrt{\log s}/\sqrt{n} \rightarrow 0. \quad (3.26)$$

Compared to (3.15), (3.26) is clearly more demanding since d_{SS}^* is always less than or equal to s . Note that a necessary condition for $\hat{\Sigma}_{SS}$ to be non-singular is $s \leq n$, which is not required for $\hat{\Sigma}_{SS}^\nu$. Thus, the non-singularity of the sample covariance sub-matrix $\hat{\Sigma}_{SS}$ is harder to attain than that of $\hat{\Sigma}_{SS}^\nu$. In other words, covariance-thresholded lasso may increase the rank of $\hat{\Sigma}_{SS}$ by thresholding. When Σ is sparse, this can be beneficial for variable selection under the large p and small n scenario.

To ensure that $P(\|(\hat{\Sigma}_{SS})^{-1}\|_\infty \leq \bar{D}) \rightarrow 1$ and $P(\|\hat{\Sigma}_{CS}(\hat{\Sigma}_{SS})^{-1}\|_\infty \leq 1 - \epsilon/2) \rightarrow 1$, as in Lemma 3.3 with $\nu = 0$, we further assume the conditions (3.17) and

$$\bar{D}^2 s^2 \sqrt{\log(p-s)}/\sqrt{n} \rightarrow 0. \quad (3.27)$$

We note that (3.27) is the main condition that guarantees that $\hat{\Sigma}$ satisfies the irrepresentable condition with probability going to 1. Compared with (3.19),

(3.27) is clearly more demanding since s is larger than both d_{CS}^* and d_{SS}^* . This implies that it is harder for $\hat{\Sigma}$ than for $\hat{\Sigma}_\nu$ to satisfy the irrepresentable condition. In other words, covariance-thresholded lasso is more likely to be variable selection consistent than the lasso when data are randomly generated from a distribution that satisfies (3.17).

Finally, with the additional condition,

$$\bar{D}\sqrt{\log s}/(\underline{\rho}\sqrt{n}) \rightarrow 0, \quad (3.28)$$

we arrive at the sign consistency of the lasso as the following.

Corollary 3.1 *Assume that the conditions (3.13), (3.14), (3.17), (3.27), and (3.28) are satisfied. If λ_n is chosen such that $\lambda_n \rightarrow 0$,*

$$\sqrt{n}\lambda_n/\sqrt{\log(p-s)} \rightarrow \infty, \quad \text{and} \quad \bar{D}\lambda_n/\underline{\rho} \rightarrow 0, \quad (3.29)$$

then, $P\left(\text{sgn}(\hat{\beta}^0) = \text{sgn}(\beta^)\right) \rightarrow 1$.*

Compare Corollary 3.1 with Theorem 3.2 for covariance-thresholded lasso. We see that conditions (3.13), (3.14), (3.17) on random predictors, in particular the covariances, are the same, but conditions on dimension parameters, such as n , p , s , etc., are different. When the population covariance matrix Σ is sparse, condition (3.19) on dimension parameters is much weaker for covariance-thresholded lasso than condition (3.27) for the lasso. This shows that covariance-thresholded lasso can improve the possibility of there existing a consistent solution. However, a trade-off presents in the selection of tuning parameters λ_n . The first condition in (3.23) for covariance-thresholded lasso is clearly more restricted than the condition in (3.29) for the lasso. This results in a more restricted range for valid λ_n . We argue that compared with the existence of consistent solution, the range of the λ_n is of secondary concern.

We note that a related sign consistency result under random design for the lasso has been established in Wainwright (2006). They assume that the predictors are normally distributed and utilize the resulting distribution of the sample covariance matrix. The conditions used in Wainwright (2006) include (3.14), (3.17), $\Lambda_{\max}(\Sigma) < \infty$, $\bar{D} < \infty$, $\log(p-s)/(n-s) \rightarrow 0$, $\sqrt{\log s}/(\underline{\rho}\sqrt{n}) \rightarrow 0$, and $n > 2\left(\Lambda_{\max}(\Sigma)/(\epsilon^2\Lambda_{\min}(\Sigma_{SS})) + \nu\right) s \log(p-s) + s + 1$, for some constant $\nu > 0$.

In comparison, we assume, in this paper, that the random predictors follow the more general moment conditions (3.13), which contain the Gaussian assumption as a special case. Moreover, we use a new approach to establish sign consistency that can incorporate the sparsity of the covariance matrix.

4. Simulations

In this section, we examine the finite-sample performances of the covariance-thresholded lasso for $p \geq n$ and compare them to those of the lasso, adaptive lasso with univariate as initial estimates, UST, scout(1,1), scout(2,1), and elastic net. Further, we propose a novel variant of cross-validation that allows improved variable selection when n is much less than p . We note that the scout(1,1) procedure can be computationally expensive. Results for scout(1,1) that take longer than 5 days on an RCAC cluster were not shown.

We compare variable selection performances using the G -measure, $G = \sqrt{\text{sensitivity} * \text{specificity}}$. G is defined as the geometric mean between sensitivity, (no. of true positives)/ s , and specificity, $1 - (\text{no. of false positives})/(p - s)$. Sensitivity and specificity can be interpreted as the proportion of selecting the true variables correctly and discarding the irrelevant variables correctly, respectively. Sensitivity can also be defined as 1 minus false negative rate and specificity as 1 minus false positive rate. A value close to 1 for G indicates good selection, whereas a value close to 0 implies that few true variables or too many irrelevant variables are selected, or both. Furthermore, we compare prediction accuracy using the relative prediction error (RPE), $\text{RPE} = (\hat{\beta} - \beta^*)^T \Sigma (\hat{\beta} - \beta^*) / \sigma^2$ where Σ is the population covariance matrix. The RPE is obtained by re-scaling the mean-squared error (ME), as in Tibshirani (1996), by $1/\sigma^2$.

We first present variable selection results using best-possible selection of tuning parameters, where tuning parameters are selected ex post facto based on the best G . This procedure is useful in examining variable selection performances, free from both inherent variabilities in estimating the tuning parameters and possible differences in the validation procedures used. Moreover, it is important as an informant of the possible potentials of the methods examined. We present median G out of 200 replications using best-possible selection of tuning parameters. Standard errors based on 500 bootstrapped re-samplings are very small, in the hundredth decimal place, for median G and are not shown.

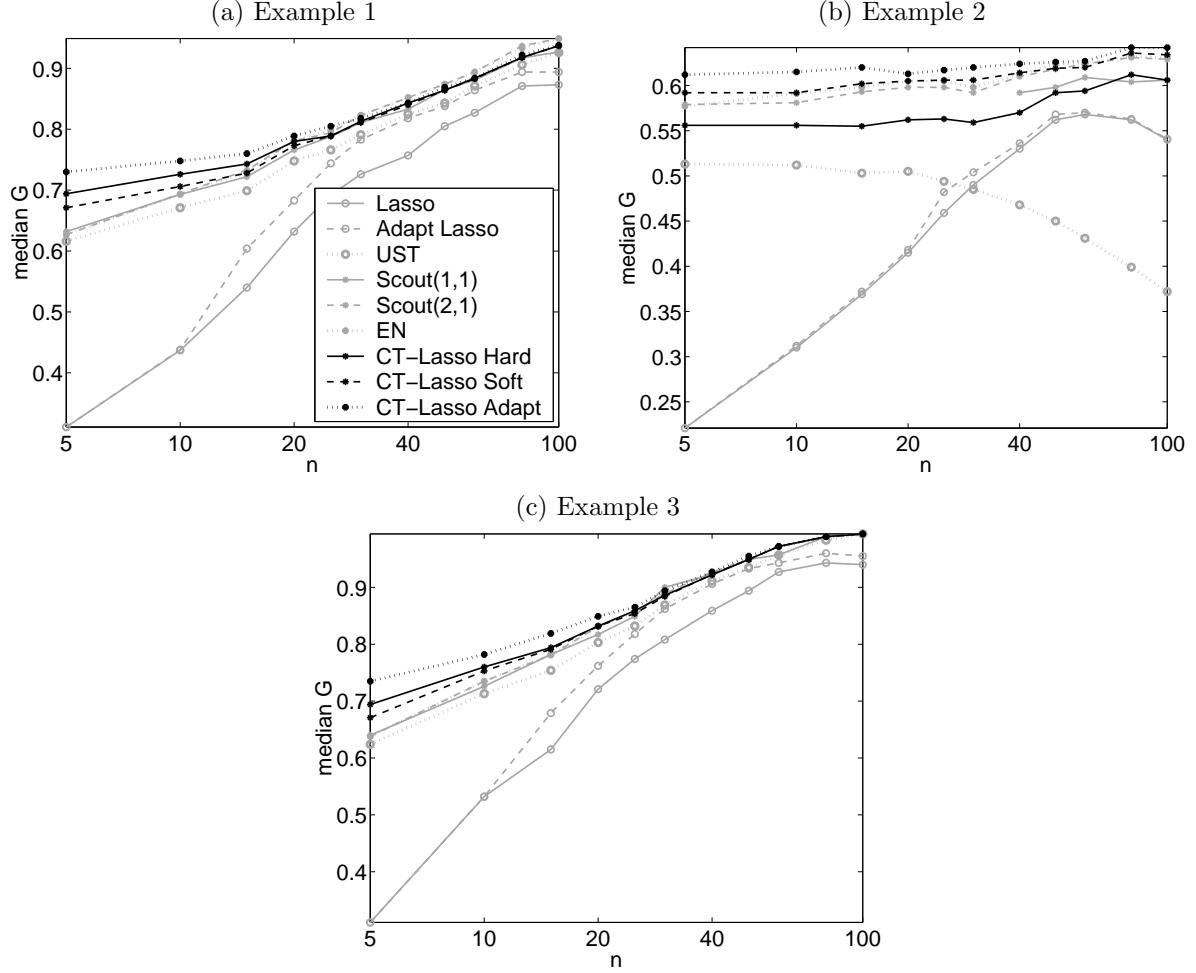


Figure 3: Variable selection performances using best-possible selection of tuning parameters based on 200 replications at $n = \{5, 10, 15, 20, 25, 30, 40, 50, 60, 80, 100\}$.

Results from best-possible selection of tuning parameters allow us to understand the potential advantages of the methods if one chooses their tuning parameters correctly. However, in practice, possible errors due to the selection of tuning parameters may sometimes overcome the benefit of introducing them. Hence, we include additional results that use cross-validation to select tuning parameters.

We study variable selection methods using a novel variant of the usual cross-validation to estimate the model complexity parameter λ_n that allows improved

variable selection when $p \gg n$. Conventional cross-validation selects tuning parameters based upon the minimum validation error, obtained from the average of sum-of-squares errors from each fold. It is well known that, when the sample size n is large compared with the number of predictors p , procedures such as cross-validation that are prediction-based tend to over-select. This is because, when the sample size is large, regression methods tend to produce small but non-zero estimates for coefficients of irrelevant variables and over-training occurs. On the other hand, we note that a different scenario occurs when $p \gg n$. In this situation, prediction-based procedures, such as the usual cross-validation, tend to under-select important variables. This is because, when n is small, inclusion of a relatively few irrelevant variables can increase the validation error dramatically, resulting in severe instability and under-representation of important variables. In this paper, we propose to use a variant of the usual cross-validation, in which we include additional variables by decreasing $\hat{\lambda}_n$ for up to 1 standard deviation of the validation error at the minimum. Through extensive empirical studies, we found that this strategy often works well to prevent under-selection when $n/\sqrt{p} < 5$, which corresponds to $n < 50$ when $p = 100$ and $n < 224$ when $p = 2000$. For $n/\sqrt{p} \geq 5$ and sample size n only moderately large, we use the usual cross-validation at the minimum. We note that Hastie, Tibshirani, and Friedman (2001, p. 216) have described a related strategy that discards variables up to 1 standard deviation of the minimum cross-validation error for use when n is large relative to p and over-selection is severe. In Table 1-3, we present median RPE, number of true and false positives, sensitivity, specificity, and G out of 200 replications using modified cross-validation for selecting tuning parameters. The smallest 3 values of median RPE and largest 3 of median G are highlighted in bold. Standard errors based on 500 bootstrapped re-samplings are further reported in parentheses for median RPE and G. In Table 4, we provide an additional simulation study to illustrate the modified cross-validation.

In each example, we simulate 200 data sets from the true model, $\mathbf{y} = \mathbf{X}\beta^* + \sigma\epsilon$, where $\epsilon \sim N(0, I)$. \mathbf{X} is generated each time from $N(0, \Sigma)$, and we vary Σ , β^* , and σ in each example to illustrate performances across a variety of situations. We choose the tuning parameter γ from $\{0, 0.5, 1, 2\}$ for both adaptive lasso

(Zou 2006) and covariance-thresholded lasso with adaptive thresholding. The adaptive lasso seeks to improve upon the lasso by applying the weights $1/|\hat{\beta}_0|^\gamma$, where $\hat{\beta}_0$ is an initial estimate, in order to penalize each coefficient differently in the $L1$ -norm of the lasso. The larger γ is the less the shrinkage applied to coefficients of large magnitudes. The candidate values used for γ are suggested in Zou (2006) and found to work well in practice.

Example 1. (Autocorrelated.) This example has $p = 100$ predictors with coefficients $\beta_j^* = 3$ for $j \in \{1, \dots, 5\}$, $\beta_j^* = 1.5$ for $j \in \{11, \dots, 15\}$, and $\beta_j^* = 0$ otherwise. $\Sigma_{ij} = 0.5^{|i-j|}$ for all i, j , and $\sigma = 9$. Signal-to-noise ratio (SNR) $\beta^{*T} \Sigma \beta^* / \sigma^2$ is approximately 1.55. This example, similar to Example 1 in (Tibshirani 1996), has an approximately sparse covariance structure, as elements away from the diagonal can be extremely small.

Figure 3(a) depicts variable selection results using best-possible selection of tuning parameters. We see that the covariance-thresholded lasso methods dominate the lasso, adaptive lasso, and UST in terms of variable selection for $p \geq n$. The performances of lasso and adaptive lasso deteriorate precipitously as n becomes small, whereas those of the covariance-thresholded lasso methods decrease at a relatively slow pace. Furthermore, the covariance-thresholded lasso methods dominate the elastic net and scout for n small. We also observe that the scout procedures and elastic net perform very similarly. This is not surprising as Witten and Tibshirani (2009) have shown in Section 2.5.1 of their paper that scout(2,1), by regularizing the inverse covariance matrix, is very similar to the elastic net.

Results from best-possible selection provide information on the potentials of the methods examined. In Table 1, we present results using cross-validation to illustrate performances in practice. The covariance-thresholded lasso methods tend to dominate the lasso, adaptive lasso, scout, and elastic net in terms of variable selection for n small. The UST presents good variable selection performances but large prediction errors. We note that, due to its large bias, the UST cannot be legitimately applied with cross-validation that uses validation error to select tuning parameters, especially when the coefficients are disparate and some correlations are large. The advantages of covariance-thresholded lasso with hard thresholding is less apparent compared with those of soft and adap-

Table 1: Example 1 performance results using fivefold cross-validation based on 200 replications.

n	Method	rpe	TP	FP	$sens$	$spec$	G
20	Lasso	1.284 (0.043)	4.0	13.0	0.40	0.86	0.577 (0.003)
	Adapt Lasso	1.301 (0.060)	4.0	12.0	0.40	0.87	0.581 (0.006)
	UST	3.001 (0.223)	7.0	28.0	0.70	0.69	0.690 (0.008)
	Scout(1,1)	1.164 (0.027)	10.0	90.0	1.00	0.00	0.000 (0.000)
	Scout(2,1)	1.474 (0.053)	6.0	39.0	0.60	0.57	0.474 (0.023)
	Elastic net	1.630 (0.097)	7.0	31.0	0.70	0.63	0.633 (0.021)
	CT-Lasso hard	1.713 (0.100)	5.0	22.5	0.60	0.77	0.593 (0.013)
	CT-Lasso soft	1.586 (0.051)	6.0	20.5	0.60	0.78	0.667 (0.007)
	CT-Lasso adapt	1.602 (0.055)	6.0	20.0	0.60	0.78	0.654 (0.006)
40	Lasso	1.095 (0.052)	6.0	27.0	0.60	0.71	0.672 (0.003)
	Adapt Lasso	1.047 (0.038)	7.0	21.0	0.70	0.77	0.706 (0.007)
	UST	1.918 (0.098)	8.0	28.0	0.80	0.69	0.742 (0.006)
	Scout(1,1)	0.814 (0.016)	10.0	90.0	1.00	0.00	0.000 (0.025)
	Scout(2,1)	1.125 (0.039)	9.0	53.0	0.90	0.41	0.544 (0.029)
	Elastic net	1.490 (0.066)	8.0	32.0	0.90	0.63	0.683 (0.010)
	CT-Lasso hard	1.221 (0.072)	7.0	23.0	0.70	0.74	0.704 (0.008)
	CT-Lasso soft	1.068 (0.055)	7.0	23.0	0.80	0.77	0.739 (0.007)
	CT-Lasso adapt	1.063 (0.045)	7.0	23.0	0.80	0.78	0.743 (0.007)
80	Lasso	0.379 (0.010)	8.0	19.0	0.80	0.79	0.794 (0.005)
	Adapt Lasso	0.367 (0.013)	8.0	15.0	0.80	0.82	0.800 (0.005)
	UST	0.360 (0.011)	8.0	5.0	0.80	0.94	0.851 (0.012)
	Scout(1,1)	0.245 (0.007)	8.0	8.0	0.80	0.91	0.854 (0.008)
	Scout(2,1)	0.399 (0.014)	6.5	7.0	0.65	0.92	0.762 (0.006)
	Elastic net	0.307 (0.014)	9.0	10.0	0.90	0.90	0.866 (0.006)
	CT-Lasso hard	0.349 (0.013)	8.0	8.0	0.80	0.94	0.795 (0.010)
	CT-Lasso soft	0.284 (0.011)	8.0	6.5	0.80	0.94	0.827 (0.008)
	CT-Lasso adapt	0.316 (0.017)	8.0	8.0	0.80	0.93	0.823 (0.009)

tive thresholding. This suggests that continuous thresholding of covariances may achieve better performances than discontinuous ones using cross-validation. We note that the scout procedures perform surprisingly poorly compared with the covariance-thresholded lasso and the elastic net in terms of variable selection when n is small. As the scout and elastic net are quite similar in terms of their potentials for variable selection as shown in Figure 3(a), the differences seem to come from the additional re-scaling step of the scout, where the scout re-scales its initial estimates by multiplying them with a scalar $\hat{c} = \arg \min_c \|\mathbf{y} - c\mathbf{X}\hat{\beta}\|^2$. This strategy can sometimes be useful in improving prediction accuracy. However, when n is small compared with p , standard deviations of validation errors for the scout can often be large, which may cause variable selection performances to suffer for cross-validation. We additionally note that, when $p \gg n$ and SNR

Table 2: Example 2 performance results using fivefold cross-validation based on 200 replications.

n	Method	rpe	TP	FP	$sens$	$spec$	G
20	Lasso	0.341 (0.027)	2.0	9.0	0.10	0.89	0.302 (0.009)
	Adapt Lasso	0.352 (0.028)	2.0	9.0	0.10	0.89	0.301 (0.006)
	Elastic net	0.967 (0.137)	14.0	51.5	0.70	0.36	0.437 (0.011)
	UST	28.930 (0.836)	19.0	73.0	0.95	0.09	0.296 (0.012)
	Scout(1,1)	NA	NA	NA	NA	NA	NA
	Scout(2,1)	0.062 (0.004)	20.0	80.0	1.00	0.00	0.000 (0.000)
	CT-Lasso hard	0.383 (0.018)	3.0	11.0	0.15	0.86	0.370 (0.013)
	CT-Lasso soft	0.231 (0.015)	6.5	23.0	0.33	0.71	0.465 (0.008)
	CT-Lasso adapt	0.302 (0.019)	6.5	23.0	0.33	0.71	0.461 (0.012)
40	Lasso	0.348 (0.017)	5.0	18.5	0.25	0.77	0.429 (0.014)
	Adapt Lasso	0.315 (0.024)	5.0	17.0	0.25	0.79	0.417 (0.008)
	Elastic net	0.739 (0.094)	16.0	58.0	0.80	0.28	0.426 (0.014)
	UST	26.189 (1.001)	20.0	77.0	1.00	0.04	0.194 (0.007)
	Scout(1,1)	NA	NA	NA	NA	NA	NA
	Scout(2,1)	0.043 (0.004)	20.0	80.0	1.00	0.00	0.000 (0.000)
	CT-Lasso hard	0.363 (0.018)	6.0	21.0	0.30	0.74	0.450 (0.008)
	CT-Lasso soft	0.269 (0.023)	10.0	35.0	0.50	0.56	0.485 (0.006)
	CT-Lasso adapt	0.306 (0.029)	8.0	31.0	0.40	0.61	0.482 (0.006)
80	Lasso	0.123 (0.004)	5.0	14.0	0.25	0.83	0.440 (0.008)
	Adapt Lasso	0.122 (0.004)	4.0	14.0	0.20	0.83	0.423 (0.009)
	Elastic net	0.089 (0.006)	14.0	48.5	0.70	0.39	0.461 (0.012)
	UST	0.042 (0.003)	18.0	66.0	0.90	0.18	0.393 (0.006)
	Scout(1,1)	NA	NA	NA	NA	NA	NA
	Scout(2,1)	0.038 (0.002)	20.0	80.0	1.00	0.00	0.000 (0.000)
	CT-Lasso hard	0.159 (0.007)	6.0	17.5	0.30	0.78	0.468 (0.008)
	CT-Lasso soft	0.107 (0.009)	9.0	27.0	0.45	0.66	0.521 (0.007)
	CT-Lasso adapt	0.129 (0.014)	8.0	24.0	0.40	0.70	0.503 (0.007)

is low as in this example, high specificity can sometimes be more important for prediction accuracy than high sensitivity. This is because, when n is small, coefficients of irrelevant variables can be given large estimates, and inclusion of but a few irrelevant variables can significantly deteriorate prediction accuracy. In Table 1, we see that the lasso and adaptive lasso have good prediction accuracy for $n = 20$ though it selects less than half of the true variables.

Example 2. (Constant covariance.) This example has $p = 100$ predictors with $\beta_j^* = 3$ for $j \in \{11, \dots, 20\}$, $\beta_j^* = 1.5$ for $j \in \{31, \dots, 40\}$, and $\beta_j^* = 0$ otherwise. $\Sigma_{ij} = 0.95$ for all i and j such that $i \neq j$, and $\sigma = 15$. SNR is approximately 8.58. This example, derived from Example 4 in Tibshirani (1996), presents an extreme situation where all non-diagonal elements of the population covariance matrix are nonzero and constant.

In Figure 3(b), we see that the covariance-thresholded lasso methods dominate over the lasso and adaptive lasso, especially for n small. This example shows that sparse covariance thresholding may still improve variable selection when the underlying covariance matrix is non-sparse. Furthermore, covariance-thresholded lasso methods with soft and adaptive thresholding perform better than that with hard thresholding. Interestingly, we see that the performance of UST decreases with increasing n and drops below that of the lasso for $n \geq 30$. This example demonstrates that the UST may not be a good general procedure for variable selection and can sometimes fail unexpectedly. We note that this is a challenging example for variable selection in general. By the irrepresentable condition (Zhao and Yu 2006), the lasso is not variable selection consistent under this scenario. The median G values in Figure 3(b) usually increase much slower with increasing n in comparison with those of Example 1 in Figure 3(a), even though SNR is higher.

Table 2 shows that the covariance-thresholded lasso methods and the elastic net dominate over the lasso and adaptive lasso in terms of variable selection when using cross-validation to select tuning parameters. The UST under-performs the lasso and adaptive lasso in terms of variable selection. Scout(2,1) does the worst in terms of variable selection by including all variables but presents the best prediction error. Again, we note that this may be due to the re-scaling step employed by the scout, which may sometimes improve performance in prediction but often suffers in terms of variable selection, especially when the sample size is small.

Example 3. (Grouped variables.) This example has $p = 100$ predictors with $\beta^* = \{3, 3, 2.5, 2.5, 2, 2, 1.5, 1.5, 1, 1, 0, \dots, 0\}$. The predictors are generated as $\mathbf{X}_j = Z_1 + \sqrt{17/3}\epsilon_{x,j}$ for $j \in \{1, \dots, 10\}$, $\mathbf{X}_j = Z_2 + \sqrt{1/19}\epsilon_{x,j}$ for $\mathbf{X}_j \in \{11, \dots, 15\}$, and $\mathbf{X}_j = \epsilon_{x,j}$ otherwise, where $Z_1 \sim N(0, 1)$, $Z_2 \sim N(0, 1)$, and $\epsilon_{x,j} \sim N(0, 1)$ are independent. This creates within-group correlations of $\Sigma_{ij} = 0.15$ for $i, j \in \{1, \dots, 10\}$ and $\Sigma_{ij} = 0.95$ for $i, j \in \{11, \dots, 15\}$. $\sigma = 15$ and SNR is approximately 1.1. This example presents an interesting scenario where a group of significant variables are mildly correlated and simultaneously a group of insignificant variables are strongly correlated.

In Figure 3(c), we see that the covariance-thresholded lasso dominates gen-

Table 3: Example 3 performance results using fivefold cross-validation based on 200 replications.

n	Method	rpe	TP	FP	$sens$	$spec$	G
20	Lasso	0.751 (0.024)	5.0	12.0	0.50	0.87	0.650 (0.003)
	Adapt Lasso	0.773 (0.035)	5.0	11.0	0.50	0.88	0.652 (0.005)
	UST	3.340 (0.202)	8.0	31.0	0.80	0.66	0.712 (0.007)
	Scout(1,1)	0.682 (0.022)	10.0	90.0	1.00	0.00	0.000 (0.000)
	Scout(2,1)	1.274 (0.081)	8.0	37.5	0.80	0.58	0.536 (0.017)
	Elastic net	1.891 (0.203)	9.0	32.0	0.90	0.64	0.660 (0.015)
	CT-Lasso hard	1.542 (0.105)	6.5	20.5	0.70	0.77	0.636 (0.018)
	CT-Lasso soft	1.240 (0.058)	7.0	22.5	0.70	0.77	0.700 (0.013)
	CT-Lasso adapt	1.427 (0.069)	7.0	19.0	0.70	0.76	0.665 (0.015)
40	Lasso	0.729 (0.044)	8.0	27.0	0.80	0.70	0.748 (0.008)
	Adapt Lasso	0.659 (0.038)	8.0	21.5	0.80	0.76	0.789 (0.008)
	UST	1.784 (0.068)	10.0	33.0	1.00	0.63	0.782 (0.010)
	Scout(1,1)	0.518 (0.013)	10.0	68.5	1.00	0.24	0.475 (0.103)
	Scout(2,1)	0.628 (0.037)	10.0	54.5	1.00	0.39	0.616 (0.028)
	Elastic net	1.101 (0.057)	10.0	35.0	1.00	0.62	0.748 (0.012)
	CT-Lasso hard	0.808 (0.045)	9.0	21.0	0.90	0.76	0.806 (0.012)
	CT-Lasso soft	0.723 (0.026)	9.0	22.0	0.90	0.76	0.815 (0.007)
	CT-Lasso adapt	0.760 (0.046)	9.0	22.0	0.90	0.76	0.819 (0.009)
80	Lasso	0.221 (0.013)	9.0	24.0	0.90	0.73	0.825 (0.007)
	Adapt Lasso	0.222 (0.017)	10.0	19.0	1.00	0.79	0.864 (0.006)
	UST	0.104 (0.008)	10.0	6.0	1.00	0.93	0.946 (0.004)
	Scout(1,1)	0.070 (0.005)	10.0	5.5	1.00	0.94	0.937 (0.004)
	Scout(2,1)	0.070 (0.003)	10.0	7.0	1.00	0.92	0.938 (0.004)
	Elastic net	0.104 (0.009)	10.0	9.0	1.00	0.90	0.937 (0.005)
	CT-Lasso hard	0.069 (0.005)	9.0	3.0	0.90	0.97	0.938 (0.004)
	CT-Lasso soft	0.063 (0.005)	10.0	4.0	1.00	0.94	0.938 (0.003)
	CT-Lasso adapt	0.063 (0.004)	10.0	4.0	1.00	0.96	0.943 (0.003)

erally in terms of variable selection. Similarly, Table 3 shows that the covariance-thresholded lasso does relatively well compared with other methods when using cross-validation to select tuning parameters. Further, the elastic net tends to have lower specificities than the covariance-thresholded lasso methods. In the related scenario of Example 4 in Zou and Hastie (2005), where a group of significant variables has strong within-group correlation and independent otherwise, the performances of elastic net are similar to those of covariance-thresholded lasso using soft thresholding, as both methods regularize covariances with large magnitudes.

Methods of Cross-Validation. We examine the modified cross-validation presented in the beginning of this section. In Table 4, we summarize results of 3 variants of cross-validation from covariance-thresholded lasso with soft threshold-

Table 4: Performances of cross-validation methods based upon 200 replications of covariance-thresholded lasso with soft thresholding. CV_- includes additional variables up to 1 standard deviation of the minimum cross-validation error; CV_0 selects λ_n at the minimum; and CV_+ discards variables up to 1 standard deviation of the minimum.

Example	n	CV_-			CV_0			CV_+		
		<i>sens</i>	<i>spec</i>	G	<i>sens</i>	<i>spec</i>	G	<i>sens</i>	<i>spec</i>	G
Ex 1	20	0.60	0.78	0.667 (0.007)	0.50	0.92	0.607 (0.009)	0.00	1.00	0.000 (0.086)
	40	0.80	0.77	0.739 (0.007)	0.60	0.92	0.699 (0.011)	0.30	1.00	0.548 (0.019)
	60	0.80	0.73	0.756 (0.013)	0.70	0.93	0.776 (0.013)	0.40	1.00	0.632 (0.016)
	80	0.90	0.73	0.789 (0.011)	0.80	0.94	0.827 (0.008)	0.50	1.00	0.707 (0.002)
Ex 2	20	0.33	0.71	0.465 (0.008)	0.23	0.83	0.409 (0.012)	0.15	0.89	0.362 (0.009)
	40	0.50	0.56	0.485 (0.006)	0.30	0.79	0.463 (0.009)	0.20	0.86	0.414 (0.014)
	60	0.55	0.54	0.504 (0.009)	0.35	0.71	0.497 (0.007)	0.30	0.79	0.473 (0.011)
	80	0.65	0.48	0.505 (0.009)	0.45	0.66	0.521 (0.007)	0.35	0.74	0.498 (0.008)
Ex 3	20	0.70	0.77	0.700 (0.013)	0.60	0.90	0.671 (0.015)	0.20	0.99	0.433 (0.117)
	40	0.90	0.76	0.815 (0.007)	0.80	0.91	0.846 (0.011)	0.60	0.99	0.762 (0.025)
	60	1.00	0.79	0.865 (0.006)	0.90	0.93	0.922 (0.004)	0.80	0.99	0.872 (0.018)
	80	1.00	0.80	0.882 (0.007)	1.00	0.94	0.938 (0.003)	0.80	1.00	0.894 (0.002)

ing. Cross-validation by including additional variables up to 1 standard deviation of the minimum (CV_-), cross-validation by minimum validation error (CV_0), and cross-validation by discarding variables up to 1 standard deviation of the minimum (CV_+) are presented. The largest G value and smallest bootstrapped standard deviations of G among cross-validation methods are highlighted in boldface.

The results demonstrate the overwhelming pattern that the proportion of relevant variables selected, or sensitivity, decreases with n under cross-validation. We note that CV_+ , as recommended in Hastie, Tibshirani, and Friedman (2001), does not work well in general for $n < p$. For n very small, CV_0 often selects too few variables, whereas, for n relatively large, CV_- usually includes too many irrelevant variables. Moreover, when n is very small, bootstrapped standard deviations of G are usually the smallest for CV_- , whereas, when n is relatively large, CV_0 usually yields better standard deviations of G . These observations suggest the modified cross-validation that employs CV_0 when $n/\sqrt{p} > 5$ and CV_- when $n/\sqrt{p} < 5$.

5. Real Data

In this section, we compare the performance of covariance-thresholded lasso with those of lasso, adaptive lasso, UST, scout(1,1), scout(2,1), and elastic net. We apply the methods to 3 well-known data sets. For each data set, we randomly

partition the data into a training and a testing set. Tuning parameters are estimated using fivefold cross-validations on the training set, and performances are measured with the testing set. When $n/\sqrt{p} < 5$, the modified cross-validation described in Section 4 is used, where additional variables are included up to 1 standard deviation of the validation error at the minimum. In order to avoid inconsistency of results due to randomization (Bøvelstad, Nygård, Størvold, Aldrin, Borgan, Frigessi, and Lingjærde 2000), we repeat the comparisons 100 times, each with a different random partition of the training and testing set. In Table 5, we report median test MSE or classification error and number of variables selected. The smallest 3 test MSEs or classification errors are highlighted in boldface. In addition, standard errors based on 500 bootstrapped re-samplings are reported in parentheses.

Highway data. Consider the highway accident data from an unpublished master’s paper by C. Hoffstedt and examined in Weisberg (1980). The data set contains 39 observations, which we divide randomly into $n = 28$ and $n_{Test} = 11$ observations for the training and testing set, respectively. The response is y =accident rate per million vehicle miles. There are originally 9 predictors, and we further include quadratic and interaction terms to obtain a total of $p = 54$ predictors. The original predictors are X_1 =length of highway segment, X_2 =average daily traffic count, X_3 =truck volume as a percentage of the total volume, X_4 =speed limit, X_5 =width of outer shoulder, X_6 =number of freeway-type interchanges per mile, X_7 =number of signalized interchanges per mile, X_8 =number of access points per mile, and X_9 =total number of lanes of traffic in both directions.

Table 5 summarizes the results obtained. Covariance-thresholded lasso methods with hard, soft, and adaptive thresholding outperform the elastic net with 12%, 44%, and 49% reductions in median tMSE, respectively, and the lasso with 21%, 49%, and 54% reductions in median tMSE, respectively. The scout has the smallest tMSE. We note that this may be due to scout’s additional re-scaling step, in which it multiplies its initial estimates by a scalar $\hat{c} = \arg \min_c \|\mathbf{y} - c\mathbf{X}\hat{\beta}\|^2$, as explained in Section 4.

CDI data. Next, we consider the county demographic information (CDI) data from the Geospatial and Statistical Data Center of the University of Virginia and examined in Kutner, Nachtsheim, Neter, and Li (2005). The data set

Table 5: Highway ($n=28$, $nTest=11$, $p=54$), CDI ($n=308$, $nTest=132$, $p=90$), and Golub microarray ($n=38$, $nTest=34$, $p=1,000$) data performance results based on 100 random partitions of training and testing sets.

<i>Method</i>	<i>Highway</i>		<i>CDI</i>		<i>Golub Microarray</i>	
	<i>tMSE</i>	<i>no.</i>	<i>tMSE/10¹⁰</i>	<i>no.</i>	<i>test error</i>	<i>no.</i>
Lasso	6.836 (0.917)	24 (0.5)	0.925 (0.225)	82 (1.8)	3.0 (0.383)	37 (0.0)
Adapt Lasso	6.246 (0.577)	22 (0.4)	0.701 (0.263)	67.5 (4.0)	3.0 (0.401)	37 (0.0)
UST	12.948 (1.138)	24 (0.9)	1.562 (0.115)	20 (0.3)	2.0 (0.447)	198 (0.0)
Scout(1,1)	3.121 (0.172)	20.5 (2.0)	NA	NA	NA	NA
Scout(2,1)	2.372 (0.292)	17.5 (1.9)	0.201 (0.014)	6 (1.1)	1.0 (0.472)	194 (2.2)
Elastic net	6.165 (0.549)	31 (2.2)	0.216 (0.013)	22.5 (1.1)	3.0 (0.336)	26.5 (5.9)
CT-Lasso hard	5.400 (0.481)	25 (1.4)	0.226 (0.021)	35.5 (4.7)	3.0 (0.388)	21 (3.7)
CT-Lasso soft	3.480 (0.268)	21 (2.5)	0.185 (0.010)	21 (2.7)	3.0 (0.388)	24.5 (3.4)
CT-Lasso adapt	3.170 (0.486)	19.5 (1.3)	0.209 (0.015)	26 (2.4)	2.5 (0.476)	36 (0.0)

contains 440 observations, which we divide randomly into $n = 308$ and $nTest = 132$ observations for the training and testing set, respectively. The response is y =total number of crimes. There are originally 12 predictors, and we further include quadratic and interaction terms to obtain a total of $p = 90$ predictors. The original predictors are X_1 =land area, X_2 =population, X_3 =percent 18-24 years old, X_4 =percent 65 years old or older, X_5 =number of active nonfederal physicians, X_6 =number of hospital beds, X_7 =percent of adults graduated from high school, X_8 =percent of adults with bachelor's degree, X_9 =percent below poverty level income, X_{10} =percent of labor force unemployed, X_{11} =per capita income, and X_{12} =total personal income.

Table 5 shows that the scout, elastic net, and covariance-thresholded lasso dominate the lasso and adaptive lasso in terms of prediction accuracy. Covariance-thresholded lasso with soft thresholding performs the best with 80% reduction in median tMSE from that of the lasso. Adaptive lasso methods with relatively large bootstrapped standard errors perform comparably to the lasso.

Microarray data. Finally, we consider the microarray data from Golub, Slonim, Tamayo, Huard, Gaasenb. This example seeks to distinguish acute leukemias arising from lymphoid precursors (ALL) and myeloid precursors (AML). The data set contains 72 observations, which we divide randomly into $n = 38$ and $nTest = 34$ observations for the training and testing set, respectively. For the response y , we assign values of 1 and -1 to ALL and AML, respectively. A classification rule is applied for the fitted response such that ALL is represented if $y \geq 0$ and AML otherwise.

There are originally 7,129 predictors from Affymetrix arrays. We use sure independence screening (SIS) with componentwise regression, as recommended in Fan and Lv (2008), to first select $p = 1,000$ candidate genes. An early stop strategy is applied for all methods at the 200th step, and cross-validation is performed using the number of steps.

Table 5 presents results in terms of test errors or the numbers of misclassifications out of 36 test samples. We note that performances of the covariance-thresholded lasso methods are comparable with those from the lasso, adaptive lasso, and elastic net in terms of prediction accuracy. However, covariance-thresholded lasso methods with hard and soft thresholding select comparably less variables than the lasso, adaptive lasso, and elastic net, whereas the scout severely over-selects with the number of variables selected close to the maximum of 200 due to early stopping. In the presence of comparable prediction accuracy, this may suggest that covariance-thresholded lasso can more readily differentiate between true and irrelevant variables under high-dimensionality.

6. Conclusion and Further Discussions

In this paper, we have proposed the covariance-thresholded lasso, a new regression method that stabilizes and improves the lasso for variable selection by utilizing covariance sparsity, which is an ubiquitous property in high-dimensional applications. The method presents as an important marriage between methods of covariance regularization (Bickel and Levina 2008a) and variable selection. We have shown theoretical studies and presented simulation and real-data examples to indicate that our method can be useful in improving variable selection performances, especially when $p \gg n$.

Furthermore, we note that there are many other variable selection procedures, such as the relaxed lasso (Meinshausen 2007), VISA (Radchenko and James 2008), etc., that may well be considered for comparison in Section 4 for the $n < p$ scenario. However, due to limit in space, we restrict ourselves to only closely related methods in this paper. We believe it can be interesting to further explore other methods for the $n < p$ scenario using modified cross-validation and best-possible selection of tuning parameters, and we hope to include them in future works.

Finally, sparse covariance-thresholding is a general procedure for variable selection in high-dimensional applications. In this paper, we applied covariance-

thresholding specifically to the lasso. Nonetheless, a myriad of variable selection methods, such as the Dantzig selector (Candes and Tao 2007), SIS (Fan and Lv 2008), etc., can also benefit by utilizing covariance-thresholding to improve variable selection. We believe that results established in this paper will also be useful in applying sparse covariance-thresholding for variable selection methods other than the lasso.

7. Appendix

In this appendix, we first state and prove some preliminary lemmas that will be used in later proofs. Lemma 7.2 gives the upper bounds of $\hat{\Sigma}_{CS}^\nu$ and $\hat{\Sigma}_{SS}^\nu$ as estimates of Σ_{CS} and Σ_{SS} , respectively. Lemma 7.3 gives the upper bound of any sample covariance matrix as an estimate of its population counterpart. The rest of the appendix is dedicated to the proofs of results in Section 3.1. The proofs of results in Section 3.2, which we omit, are similar to those in Section 3.1, except that ν is set to be 0 and Lemma 7.3 is used in place of Lemma 7.2.

7.1. Preliminary Lemmas

Lemma 7.1 *Suppose $(X_{k1}, X_{k2}, \dots, X_{kp})$, $1 \leq k \leq n$, are independent and identically distributed random vectors with $E(X_{kj}) = 0$, $E(X_{ki}X_{kj}) = \sigma_{ij}$, and $EX_{kj}^{2d} \leq d!M^d$ for $d \in \mathbb{N} \cup \{0\}$, $M > 0$ and $1 \leq i, j \leq p$. Let $\hat{\sigma}_{ij} = \frac{1}{n} \sum_{k=1}^n X_{ki}X_{kj}$. Then, for $t_n = o(1)$,*

$$P(|\hat{\sigma}_{ij} - \sigma_{ij}| > t_n) \leq \exp(-cnt_n^2), \quad (7.30)$$

where c is some constant depending only on M .

Proof of Lemma 7.1

Let $Z_k = X_{ki}X_{kj} - \sigma_{ij}$. We apply the Bernstein's Inequality (moment version) (see for example van der Vaart and Wellner (1996)) on the series $\sum_{k=1}^n Z_k$.

For $m \geq 1$, we have $E|Z_k|^m = E|X_{ki}X_{kj} - \sigma_{ij}|^m \leq \sum_{d=0}^m \binom{m}{d} |\sigma_{ij}|^{m-d} E|X_{ki}X_{kj}|^d$. By the moment conditions in Lemma 7.1, we have $|\sigma_{ij}| \leq M$ and $E|X_{ki}X_{kj}|^d \leq \frac{1}{2} (EX_{ki}^{2d} + EX_{kj}^{2d}) \leq d!M^d$. Therefore, $E|Z_k|^m \leq m!M^m \sum_{d=0}^m \binom{m}{d} = m!(2M)^m$, and result follows by applying the moment version of Bernstein's Inequality. \square

Lemma 7.2 *If ν is chosen to be greater than $C\sqrt{\log(s(p-s))}/\sqrt{n}$ for some C large enough, then*

$$\left\| \hat{\Sigma}_{CS}^\nu - \Sigma_{CS} \right\|_\infty \leq O_p(\nu d_{CS}^*) + O_p\left(d_{CS}^* \sqrt{\log(s(p-s))}/\sqrt{n}\right). \quad (7.31)$$

If ν is chosen to be greater than $C\sqrt{\log s}/\sqrt{n}$ for some C large enough, then

$$\left\| \hat{\Sigma}_{SS}^\nu - \Sigma_{SS} \right\|_\infty \leq O_p(\nu d_{SS}^*) + O_p\left(d_{SS}^* \sqrt{2\log s}/\sqrt{n}\right). \quad (7.32)$$

The proof is similar to that of Theorem 1 in Bickel and Levina (2008a) and Theorem 1 in Rothman et al. (2009), and, thus, it is omitted to save space. The detailed proof can be found in the supplementary document.

Lemma 7.3 *Let A and B be two arbitrary subsets of $\{1, 2, \dots, p\}$, and let $\Sigma_{AB} = (\sigma_{ij})_{i \in A, j \in B}$ and $\hat{\Sigma}_{AB} = (\hat{\sigma}_{ij})_{i \in A, j \in B}$. Further, let a be the cardinality of A and b the cardinality of B . Suppose a and b satisfy $\sqrt{\log(ab)}/\sqrt{n} \rightarrow 0$ as $n \rightarrow \infty$. Then*

$$\left\| \hat{\Sigma}_{AB} - \Sigma_{AB} \right\|_\infty = O_p\left(b\sqrt{\log(ab)}/\sqrt{n}\right).$$

Proof of Lemma 7.3

Since

$$P\left(\left\| \hat{\Sigma}_{AB} - \Sigma_{AB} \right\|_\infty > t\right) \leq \sum_{i \in A} \sum_{j \in B} P(|\hat{\sigma}_{ij} - \sigma_{ij}| > t/b) \leq a \cdot b \cdot \exp(-cnt^2/b^2),$$

for $t/b = o(1)$ by Lemma 7.1, the result follows. \square

7.2. Proof of Lemma 3.1

By the KKT conditions, the solution of (2.3) satisfies $\hat{\Sigma}_\nu \hat{\beta}^\nu - \frac{1}{n} \mathbf{X}^T y + \lambda_n \hat{z} = 0$, where \hat{z} is the sub-gradient of $\|\hat{\beta}^\nu\|_1$, that is, $\hat{z} = \partial \|\hat{\beta}^\nu\|_1$. Plugging in $y = \mathbf{X}\beta^* + \epsilon$, we have

$$\hat{\Sigma}_\nu(\hat{\beta}^\nu - \beta^*) + (\hat{\Sigma}_\nu - \hat{\Sigma})\beta^* - \frac{1}{n} \mathbf{X}^T \epsilon + \lambda_n \hat{z} = 0. \quad (7.33)$$

It is easy to see that $\text{sgn}(\hat{\beta}^\nu) = \text{sgn}(\beta^*)$ holds if $\hat{\beta}_S^\nu \neq 0$, $\hat{\beta}_C^\nu = 0$, $\hat{z}_S = \text{sgn}(\beta_S^*)$, and $|\hat{z}_C| \leq 1$. Therefore, based on (7.33), the conditions for $\text{sgn}(\hat{\beta}^\nu) = \text{sgn}(\hat{\beta}^*)$ to hold are

$$\hat{\Sigma}_{SS}^\nu(\hat{\beta}_S^\nu - \beta_S^*) + (\hat{\Sigma}_{SS}^\nu - \hat{\Sigma}_{SS})\beta_S^* - \frac{1}{n} \mathbf{X}_S^T \epsilon = -\lambda_n \text{sgn}(\beta_S^*), \quad (7.34)$$

$$\text{sgn}(\hat{\beta}_S^\nu) = \text{sgn}(\beta_S^*), \quad (7.35)$$

$$\left\| \hat{\Sigma}_{CS}^\nu (\hat{\beta}_S^\nu - \beta_S^*) + (\hat{\Sigma}_{CS}^\nu - \hat{\Sigma}_{CS}) \beta_S^* - \frac{1}{n} \mathbf{X}_C^T \epsilon \right\|_\infty \leq \lambda_n. \quad (7.36)$$

Solving (7.34) for $\hat{\beta}_S^\nu$ under the assumption $\Lambda_{\min}(\hat{\Sigma}_{SS}^\nu) > 0$, we have

$$\hat{\beta}_S^\nu = \beta_S^* + (\hat{\Sigma}_{SS}^\nu)^{-1} \left(\frac{1}{n} \mathbf{X}_S^T \epsilon - \lambda_n \text{sgn}(\beta_S^*) - (\hat{\Sigma}_{SS}^\nu - \hat{\Sigma}_{SS}) \beta_S^* \right). \quad (7.37)$$

Substituting (7.37) into the left-hand side of (7.36) and further decomposing the resulting equation, we have

$$\begin{aligned} & \left\| \hat{\Sigma}_{CS}^\nu (\hat{\Sigma}_{SS}^\nu)^{-1} \left(\frac{1}{n} \mathbf{X}_S^T \epsilon - \lambda_n \text{sgn}(\beta_S^*) - (\hat{\Sigma}_{SS}^\nu - \hat{\Sigma}_{SS}) \beta_S^* \right) + (\hat{\Sigma}_{CS}^\nu - \hat{\Sigma}_{CS}) \beta_S^* - \frac{1}{n} \mathbf{X}_C^T \epsilon \right\|_\infty \\ & \leq \left\| \hat{\Sigma}_{CS}^\nu (\hat{\Sigma}_{SS}^\nu)^{-1} \right\|_\infty \left(\left\| \frac{1}{n} \mathbf{X}_S^T \epsilon \right\|_\infty + \lambda_n + \left\| (\hat{\Sigma}_{SS}^\nu - \hat{\Sigma}_{SS}) \beta_S^* \right\|_\infty \right) \\ & + \left\| (\hat{\Sigma}_{CS}^\nu - \hat{\Sigma}_{CS}) \beta_S^* \right\|_\infty + \left\| \frac{1}{n} \mathbf{X}_C^T \epsilon \right\|_\infty \\ & \leq \left\| \hat{\Sigma}_{CS}^\nu (\hat{\Sigma}_{SS}^\nu)^{-1} \right\|_\infty \left(\left\| \frac{1}{n} \mathbf{X}_S^T \epsilon \right\|_\infty + s\nu\bar{\rho} + \lambda_n \right) + s\nu\bar{\rho} + \left\| \frac{1}{n} \mathbf{X}_C^T \epsilon \right\|_\infty, \end{aligned}$$

where the last inequality is obtained by

$$\left\| (\hat{\Sigma}_{SS}^\nu - \hat{\Sigma}_{SS}) \beta_S^* \right\|_\infty \leq \left\| \hat{\Sigma}_{SS}^\nu - \hat{\Sigma}_{SS} \right\|_\infty \|\beta_S^*\|_\infty \leq s\nu\bar{\rho}, \quad (7.38)$$

$$\left\| (\hat{\Sigma}_{CS}^\nu - \hat{\Sigma}_{CS}) \beta_S^* \right\|_\infty \leq \left\| \hat{\Sigma}_{CS}^\nu - \hat{\Sigma}_{CS} \right\|_\infty \|\beta_S^*\|_\infty \leq s\nu\bar{\rho}.$$

Then, condition (3.11) is sufficient for (7.36) to hold.

Next, we derive (3.12). By (7.37), (7.35) is implied by

$$\left\| (\hat{\Sigma}_{SS}^\nu)^{-1} \right\|_\infty \left(\left\| (\hat{\Sigma}_{SS}^\nu - \hat{\Sigma}_{SS}) \beta_S^* \right\|_\infty + \left\| \frac{1}{n} \mathbf{X}_S^T \epsilon \right\|_\infty + \lambda_n \right) < \underline{\rho}. \quad (7.39)$$

Plugging in the upper bound of $\|(\hat{\Sigma}_{SS}^\nu - \hat{\Sigma}_{SS})\beta_S^*\|_\infty$ in (7.38), it is straightforward to see that (3.12) is sufficient for (7.35) to hold. \square

7.3. Proof of Lemma 3.2

For any v with $\|v\| = 1$,

$$v^T \hat{\Sigma}_{SS}^\nu v \geq \Lambda_{\min}(\Sigma_{SS}) - \|\hat{\Sigma}_{SS}^\nu - \Sigma_{SS}\| \geq \Lambda_{\min}(\Sigma_{SS}) - \|\hat{\Sigma}_{SS}^\nu - \Sigma_{SS}\|_\infty,$$

and, when choosing $\nu = C\sqrt{\log s}/\sqrt{n}$ for some $C > 0$,

$$\|\hat{\Sigma}_{SS}^\nu - \Sigma_{SS}\|_\infty \leq O_p\left(d_{SS}^* \sqrt{\log s}/\sqrt{n}\right) \quad (7.40)$$

by Lemma 7.2. Therefore, the result follows under the condition (3.15). \square

7.4. Proof of Lemma 3.3

To derive the upper bound of $\|(\hat{\Sigma}_{SS}^\nu)^{-1}\|_\infty$, we perform the following decomposition,

$$\left\|(\hat{\Sigma}_{SS}^\nu)^{-1}\right\|_\infty \leq \left\|(\Sigma_{SS})^{-1}\right\|_\infty + \left\|(\hat{\Sigma}_{SS}^\nu)^{-1} - (\Sigma_{SS})^{-1}\right\|_\infty. \quad (7.41)$$

Because

$$\begin{aligned} \left\|(\hat{\Sigma}_{SS}^\nu)^{-1} - (\Sigma_{SS})^{-1}\right\|_\infty &\leq \left\|(\Sigma_{SS})^{-1}\right\|_\infty \left\|(\hat{\Sigma}_{SS}^\nu)^{-1}\right\|_\infty \left\|\hat{\Sigma}_{SS}^\nu - \Sigma_{SS}\right\|_\infty \\ &\leq \bar{D} \left(\left\|(\Sigma_{SS})^{-1}\right\|_\infty + \left\|(\hat{\Sigma}_{SS}^\nu)^{-1} - (\Sigma_{SS})^{-1}\right\|_\infty \right) \left\|\hat{\Sigma}_{SS}^\nu - \Sigma_{SS}\right\|_\infty \\ &= \bar{D}^2 \left\|\hat{\Sigma}_{SS}^\nu - \Sigma_{SS}\right\|_\infty + \bar{D} \left\|(\hat{\Sigma}_{SS}^\nu)^{-1} - (\Sigma_{SS})^{-1}\right\|_\infty \left\|\hat{\Sigma}_{SS}^\nu - \Sigma_{SS}\right\|_\infty, \end{aligned}$$

where the second inequality is obtained by (7.41), we have

$$\left\|(\hat{\Sigma}_{SS}^\nu)^{-1} - (\Sigma_{SS})^{-1}\right\|_\infty \leq \frac{\bar{D}^2 \left\|\hat{\Sigma}_{SS}^\nu - \Sigma_{SS}\right\|_\infty}{1 - \bar{D} \left\|\hat{\Sigma}_{SS}^\nu - \Sigma_{SS}\right\|_\infty} \leq O_p\left(\frac{\bar{D}^2 d_{SS}^* \sqrt{\log(p-s)}}{\sqrt{n}}\right) \quad (7.42)$$

where the last inequality is derived by choosing $\nu = C\sqrt{\log(p-s)}/\sqrt{n}$, applying (7.32) in Lemma 7.2, and using the condition (3.18). Combining (7.41), (7.42), and condition (3.18), we have

$$\left\|(\hat{\Sigma}_{SS}^\nu)^{-1}\right\|_\infty \leq O_p(\bar{D}). \quad (7.43)$$

For (3.21), we decompose $\hat{\Sigma}_{CS}^\nu (\hat{\Sigma}_{SS}^\nu)^{-1}$ into three terms as follows:

$$\begin{aligned} \hat{\Sigma}_{CS}^\nu (\hat{\Sigma}_{SS}^\nu)^{-1} &= \hat{\Sigma}_{CS}^\nu \left[(\hat{\Sigma}_{SS}^\nu)^{-1} - (\Sigma_{SS})^{-1} \right] + \left[\hat{\Sigma}_{CS}^\nu - \Sigma_{CS} \right] (\Sigma_{SS})^{-1} + \Sigma_{CS} (\Sigma_{SS})^{-1} \\ &= \text{I} + \text{II} + \text{III}. \end{aligned}$$

By condition (3.17), $\|\text{III}\|_\infty \leq 1 - \epsilon$. Therefore, it is enough to show that $\|\text{I}\|_\infty + \|\text{II}\|_\infty \leq \epsilon/2$ with probability going to 1.

For $\|\mathbf{II}\|_\infty$, we have

$$\|\mathbf{II}\|_\infty \leq \|(\boldsymbol{\Sigma}_{SS})^{-1}\|_\infty \|\hat{\boldsymbol{\Sigma}}_{CS}^\nu - \boldsymbol{\Sigma}_{CS}\|_\infty = \bar{D} \cdot \|\hat{\boldsymbol{\Sigma}}_{CS}^\nu - \boldsymbol{\Sigma}_{CS}\|_\infty, \quad (7.44)$$

and, when choosing $\nu = C\sqrt{\log(s(p-s))}/\sqrt{n}$,

$$\|\hat{\boldsymbol{\Sigma}}_{CS}^\nu - \boldsymbol{\Sigma}_{CS}\|_\infty \leq O_p\left(d_{CS}^* \sqrt{\log(s(p-s))}/\sqrt{n}\right) \quad (7.45)$$

by Lemma 7.2. For $\|\mathbf{I}\|_\infty$, we have

$$\begin{aligned} \|\mathbf{I}\|_\infty &\leq \|\hat{\boldsymbol{\Sigma}}_{CS}^\nu\|_\infty \left\| \left(\hat{\boldsymbol{\Sigma}}_{SS}^\nu \right)^{-1} - (\boldsymbol{\Sigma}_{SS})^{-1} \right\|_\infty \\ &\leq \left(\|\boldsymbol{\Sigma}_{CS}\|_\infty + \|\hat{\boldsymbol{\Sigma}}_{CS}^\nu - \boldsymbol{\Sigma}_{CS}\|_\infty \right) \left\| \left(\hat{\boldsymbol{\Sigma}}_{SS}^\nu \right)^{-1} - (\boldsymbol{\Sigma}_{SS})^{-1} \right\|_\infty \\ &\leq O_p\left(\bar{D}^2 d_{CS}^* d_{SS}^* \sqrt{\log(s(p-s))}/\sqrt{n}\right) \end{aligned} \quad (7.46)$$

by (7.45), (7.42), and (3.19).

In summary, we have $P(\|\mathbf{I}\|_\infty + \|\mathbf{II}\|_\infty \leq \epsilon/2) \rightarrow 1$ under the condition (3.19). This completes the proof. \square

7.5. Proof of Theorem 3.2

First, we consider $\left\| \frac{1}{n} \mathbf{X}_C^T \epsilon \right\|_\infty$ and $\left\| \frac{1}{n} \mathbf{X}_S^T \epsilon \right\|_\infty$, which appear in (3.11) and (3.12), respectively. Since $\epsilon \sim N(0, \sigma^2)$, then, when \mathbf{X} is fixed, by standard results on the extreme value of multivariate normal, we have

$$\left\| \frac{1}{n} \mathbf{X}_C^T \epsilon \right\|_\infty = O_p\left(\sigma \sqrt{2(\max_j \hat{\sigma}_{jj}) \log(p-s)}/\sqrt{n}\right), \quad (7.47)$$

$$\left\| \frac{1}{n} \mathbf{X}_S^T \epsilon \right\|_\infty = O_p\left(\sqrt{2(\max_j \hat{\sigma}_{jj}) \log s}/\sqrt{n}\right). \quad (7.48)$$

By Lemma 7.1,

$$P\left(\max_j \hat{\sigma}_{jj} > M + t\right) \leq \sum_{j=1}^p P(\hat{\sigma}_{jj} > M + t) \leq \sum_{j=1}^p P(\hat{\sigma}_{jj} - \sigma_{jj} > t) \leq p \cdot \exp(-nt^2/4)$$

for $t = o(1)$, and, thus, $P(\max_j \hat{\sigma}_{jj} \leq M) \rightarrow 1$. Therefore,

$$\left\| \frac{1}{n} \mathbf{X}_C^T \epsilon \right\|_\infty = O_p\left(\sqrt{\log(p-s)}/\sqrt{n}\right), \quad \left\| \frac{1}{n} \mathbf{X}_S^T \epsilon \right\|_\infty = O_p\left(\sqrt{\log s}/\sqrt{n}\right). \quad (7.49)$$

Now, we sum up the results in Lemma 3.1, 3.2, 3.3, and (7.49). Under conditions (3.13), (3.14), (3.17), (3.19), (3.22), and the choice of $\nu = C\sqrt{\log(p-s)}/\sqrt{n}$ for some C and λ_n as in (3.23), both (3.11) and (3.12) hold with probability going to 1. \square

7.6. Outline of the Proof of Theorem 2.1

To circumvent the problem of having a non-differentiable penalty function, we reformulate the optimization problem in (2.3) as the following,

$$\begin{aligned} \arg \min_{\beta^+, \beta^-} \quad & \frac{1}{2}(\beta^+ - \beta^-)^T \hat{\Sigma}_\nu (\beta^+ - \beta^-) - (\beta^+ - \beta^-)^T \left(\frac{1}{n} \mathbf{X}^T \mathbf{y} \right), \\ \text{s. t.} \quad & \beta_j^- \geq 0 \forall j, \quad \beta_j^+ \geq 0 \forall j, \quad \sum_j (\beta_j^+ + \beta_j^-) \leq t. \end{aligned}$$

Consider the Lagrangian primal function for the above formulation,

$$\frac{1}{2}(\beta^+ - \beta^-)^T \hat{\Sigma}_\nu (\beta^+ - \beta^-) - (\beta^+ - \beta^-)^T \left(\frac{1}{n} \mathbf{X}^T \mathbf{y} \right) + \lambda \sum_{j=1}^p (\beta_j^+ + \beta_j^-) - \sum_{j=1}^p \lambda_j^+ \beta_j^+ - \sum_{j=1}^p \lambda_j^- \beta_j^-.$$

Let $\beta = \beta^+ - \beta^-$. We obtain the following first-order conditions,

$$\begin{aligned} \frac{1}{n} \mathbf{x}_j^T \mathbf{y} - (\hat{\Sigma}_\nu)_j^T \beta - \lambda + \lambda_j^+ &= 0, \quad \frac{1}{n} \mathbf{x}_j^T \mathbf{y} - (\hat{\Sigma}_\nu)_j^T \beta + \lambda - \lambda_j^- = 0, \\ \lambda_j^+ \beta_j^+ &= 0, \quad \lambda_j^- \beta_j^- = 0. \end{aligned}$$

These conditions can be verified, as in (Rosset and Zhu 2007), to imply the facts,

$$|\frac{1}{n} \mathbf{x}_j^T \mathbf{y} - (\hat{\Sigma}_\nu)_j^T \beta| < \lambda \implies \beta_j = 0 \quad \text{and} \quad \beta_j \neq 0 \implies |\frac{1}{n} \mathbf{x}_j^T \mathbf{y} - (\hat{\Sigma}_\nu)_j^T \beta| = \lambda.$$

When $\hat{\Sigma}_\nu$ is semi-positive definite, first-order conditions are enough to provide a global solution, which is unique if all eigenvalues are positive. However, when there exist eigenvalues of $\hat{\Sigma}_\nu$ that are negative, a second-order condition, in addition to first-order ones, is required to guarantee that a point β is a local minimum. Assume strict complementarity $\beta_j = 0 \implies \lambda_j^+ > 0$ and $\lambda_j^- > 0$, which holds with high probability as regression methods rarely yield zero-valued coefficient estimate without penalization. We see that $\mathcal{K} = \{z \doteq z^+ - z^- \neq 0 : z_j^+ = 0 \text{ and } z_j^- = 0 \text{ for } \beta_j = 0\}$ covers the set of feasible directions in Theorem 6, McCormick (1976). Let $\mathcal{A} = \{j : \beta_j \neq 0\}$. By Theorem 6, McCormick (1976), a solution β is a local minimum if for every $z \in \mathcal{K}$

$$z^T (\hat{\Sigma}_\nu) z = (z_{\mathcal{A}})^T (\hat{\Sigma}_\nu)_{\mathcal{A}} z_{\mathcal{A}} > 0.$$

Furthermore, we note that the solution β is global if $|\mathbf{x}_j^T \mathbf{y}/n| < \lambda$ for all $j \notin \mathcal{A}$ in addition to $(\hat{\Sigma}_\nu)_{\mathcal{A}}$ being positive definite. This follows from facts implied by first-order conditions.

Algorithm for computing piecewise-linear solutions for the covariance-thresholded lasso is derived by further manipulating the first-order conditions as in the proof for Theorem 2 in Rosset and Zhu (2007).

Acknowledgment

The authors are grateful to Jayanta K. Ghosh and Jian Zhang for helpful comments and discussions. Furthermore, we thank the associate editor and two referees who have been very generous in providing us with helpful suggestions. Z. John Daye is supported by Purdue Research Foundation Fellowship. Computing resources and support were provided by the Department of Statistics, Purdue University, and the Rosen Center for Advanced Computing (RCAC) of Information Technology at Purdue.

References

- Bickel, P. J. and E. Levina (2008a). Covariance regularization by thresholding. *Ann. Statist.* **36**, 2577-2604.
- Bickel, P. J. and E. Levina (2008b). Regularized estimation of large covariance matrices. *Ann. Statist.* **36**, 199-227.
- Bøvelstad, H. M., S. Nygård, H. L. Størvold, M. Aldrin, Ø. Borgan, A. Frigessi, and O. C. Lingjærde (2007). Predicting survival from microarray data-a comparative study. *Bioinformatics* **23**, 2080-2087.
- Candes, E. and T. Tao (2007). The dantzig selector: Statistical estimation when p is much larger than n . *Ann. Statist.* **35**, 2313-2351.
- Chong, I. and C. Jun (2005). Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems* **78**, 103-112.
- Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *Ann. Statist.* **32**, 407-499.

- El Karoui, N. (2008). Operator norm consistent estimation of large dimensional sparse covariance matrices. *Ann. Statist.* **36**, 2717-2756.
- Fan, J. and J. Lv (2008). Sure independence screening for ultra-high dimensional feature space. *J. R. Statist. Soc. B* **70**, 849-911.
- Furrer, R. and T. Bengtsson (2007). Estimation of high-dimensional prior and posterior covariance matrices in kalman filter variants. *Journal of Multivariate Analysis* **98**, 227-255.
- Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-537.
- Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning*. Springer.
- Kubat, M., R. C. Holte, and S. Matwin (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine Learning* **30**, 195-215.
- Kutner, M. H., C. J. Nachtsheim, J. Neter, and W. Li (2005). *Applied Linear Statistical Models* (5th ed.). McGraw-Hill/Irwin.
- McCormick, G. P. (1976). Optimality criteria in nonlinear programming. In R. W. Cottle and C. E. Lemke (Eds.), *Nonlinear Programming*, Volume IX, pp. 27-38. SIAM-AMS Proceedings.
- Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics and Data Analysis* **52**, 374-393.
- Meinshausen, N. and P. Bühlmann (2006). High dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34**, 1436-1462.
- Radchenko, P. and G. James (2008). Variable inclusion and shrinkage algorithms. *J. Am. Statist. Ass.* **103**, 1304-1315.
- Rosset, S. and J. Zhu (2007). Piecewise linear regularized solution. *Ann. Statist.* **35**, 1012-1030.

- Rothman, A. J., E. Levina, and J. Zhu (2009). Generalized thresholding of large covariance matrices. *J. Am. Statist. Ass.* **104**, 177-186.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B* **58**, 267-288.
- Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005). Sparsity and smoothness via the fused lasso. *J. R. Statist. Soc. B* **67**, 91-108.
- van der Vaart, A. W. and J. A. Wellner (1996). *Weak convergence and empirical processes*. Springer.
- Wagaman, A. S. and E. Levina (2008). Discovering sparse covariance structure with the isomap. *Journal of Computational and Graphical Statistics (To appear)*.
- Wainwright, M. J. (2006). Sharp thresholding for high-dimensional and noisy recovery of sparsity. Technical Report 709, Dept. of Statistics and Department of Electrical Engineering and Computer Sciences, Univeristy of California, Berkeley.
- Weisberg, S. (1980). *Applied Linear Regression*. John Wiley & Sons.
- Witten, D. M. and R. Tibshirani (2009). Covariance-regularized regression and classification for high-dimensional problems. *J. R. Statist. Soc. B* **71**.
- Zhao, P. and B. Yu (2006). On model selection consistency of lasso. *Journal of Machine Learning Research* **7**, 2541-2567.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Am. Statist. Ass.* **101**, 1418-1429.
- Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B* **67**, 301-320.

Department of Statistics, Purdue University, West Lafayette, IN 47906, U.S.A.

E-mail: xingejeng@gmail.com

Department of Statistics, Purdue University, West Lafayette, IN 47906, U.S.A.

E-mail: zhongyindaye@gmail.com